Special issue: comparative survey analysis – comparability and equivalence of measures

Meuleman, Bart ; Davidov, Eldad ; Seddig, Daniel

mda
methods, data, analyses
Journal for Quantitative Methods and Survey Methodology

Volume 12, 2018 | 1

Comparative Survey Analysis: Comparability and Equivalence of Measures

Bart Meuleman, Eldad Davidov, & Daniel Seddig (Editors)

Edited by Annelies G. Blom, Edith de Leeuw, Gabriele Durrant

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246282
E-mail: mda@gesis.org
Internet: www.mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession's knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.

All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Print: Bonifatius Druck GmbH Paderborn, Germany

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

Content

methods, data, analyses | Vol. 12(1), 2018, pp. 3-6

Editorial: Comparative Survey Analysis – Comparability and Equivalence of Measures

Bart Meuleman[1], *Eldad Davidov*[2] *& Daniel Seddig*[2]

[1] *KU Leuven*
[2] *University of Cologne, and University of Zurich*

Over the last decades, the increasing availability of comparative survey data has opened up a wide avenue of research opportunities for social scientists. International survey projects -such as the European Social Survey (ESS), the European (EVS) and World Values Studies (WVS), or the European Household Panel Study (EHPS)- measure a wide range of attitudes and behaviors with the explicit purpose of making comparisons across countries, regions or time points (Lynn, Japec, & Lyberg, 2005). The potential relevance of such comparisons is paramount. Besides identifying differences between contexts and cultures, comparative data is helpful in testing theories about social change and contextual influences on individual characteristics. The insight that comparison is a crucial methodological tool is not new, but is as old as social science itself. After all, Durkheim (1964, p. 139) already argued that "*comparative sociology is not a particular branch of sociology: it is sociology itself*".

The advantages of the comparative design come at a methodological price, however. Collecting and analyzing cross-national survey data brings along additional methodological challenges (Berry et al., 1992; Harkness et al., 2003; Harkness et al., 2010; van de Vijver & Leung 1997). Among a great many pressing methodological issues, comparative research hinges crucially on the assumption that measurements are comparable or equivalent (Horn & McArdle, 1992; Johnson 1998; Davidov et al., 2014). Respondents in international surveys were socialized in different cultural backgrounds, speak different languages and have cultural-specific understandings of certain ideas and concepts. Therefore, it is not guaranteed that survey measurements travel successfully across national and cultural borders (Jowell et al., 2007). Equally important is to guarantee that measurements travel successfully across groups within countries (Davidov & Siegers, 2010; Sarrasin, Green, Berchtold & Davidov, 2012), across modes of data collection (Cieciuch & Davidov, 2016), or across time (Widaman, Ferrer, & Conger, 2010). Therefore, the

validity of comparisons of survey measurements across groups and time is of great concern (Jowell, 1998).

Fortunately, in recent years comparative researchers have increasingly acknowledged the importance of the comparability of measurements. A variety of methodologies have been proposed to assess to what extent survey measurements are cross-culturally equivalent (Davidov, Schmidt, Billiet & Meuleman, 2018). This special issue has the ambition to contribute to the contemporary debates on the comparability of survey measures. By providing new tools, novel insights and original applications in the field of measurement equivalence, this collection of papers advances our current knowledge on measurement equivalence.

A first set of three papers shows how measurement equivalence of multiple-item scales can be tested using a multiple-group factor analytic approach. *Wiebke Breustedt* argues that the generalizability of theories on political trust requires that this concept should be measured in a comparable way. Analyzing data from various rounds of the WVS by means of multiple group confirmatory factor analysis (MGCFA), Breustedt shows that this assumption should indeed not be taken for granted: Only in 19 out of 32 investigated democracies, configural invariance could be established. This important finding calls for a further development of cross-culturally robust instruments to gauge citizens' trust in public institutions. *Maksim Rudnev* and colleagues extend the popular MGCFA equivalence test to higher-order factor models. This paper explains in detail which model constraints are necessary to operationalize various levels of equivalence in second-order factor models. In addition, an empirical illustration evaluating the equivalence of Seeman's second-order concept of alienation across eight countries is provided. The study by *Vera Lomazzi* addresses an important weakness in the MGCFA strategy, namely that the requirements for equivalence are very strict, especially when a comparison involves a large number of groups. Lomazzi proposes to use the recently introduced alignment optimization procedure as an alternative for the common MGCFA model. Analyzing the gender role attitudes scale in the WVS across 59 countries, the results indicate that the alignment procedure is less strict and suggests that valid comparisons are possible across a wider range of countries than when the classical MGCFA model is used.

Two papers investigate how particularities of languages and writing might affect cross-cultural comparability. *Dagmar Krebs* and *Yaacov G. Bachner* tackle the intriguing question how the direction of writing – left to right vs. right to left - interacts with the way in which respondents use response scales. After all, respondents can pick up information from response scales (incremental or decremental) and factor this in their response behavior. To test this expectation, the authors analyze data from a split-ballot design among German and Israeli students. The results indicate that clear response-order effects are present, but that they are very similar in left-to-right (German) and right-to-left (Hebrew) reading directions.

Diana Zavala-Rojas studies if the language in which a survey was conducted has a noticeable impact on measurements of various political attitudes among bilingual citizens. Concretely, bilingual respondents' institutional trust and satisfaction with politics and economy were measured twice in a different language. Within-subject equivalence tests show measurements are largely equivalent across the language of survey administration, even if the correlation between two language-versions of a latent variable is not identical to 1. Summarizing, the message of these two papers is optimistic: If the necessary precautions are taken, characteristics of languages are not insurmountable for comparative researchers.

Finally, the paper by *Silke L. Schneider* draws our attention to the important message that equivalence not only matters for subjective concepts measured by multiple items. Also objective social-structural characteristics, such as educational attainment, need to be measured in a comparable way. Schneider assesses the comparability of the education variable included in PIAAC (Programme for the International Assessment of Adult Competencies). Equivalence is evaluated from the perspective of construct validity, that is, by looking at the relationship with respondents' general skills. The study shows that especially decisions to collapse the detailed education variable into a smaller number of categories challenge comparability, and identifies several pitfalls in the educational attainment variables currently used in comparative research (such as the lack of differentiation between general and vocational training).

References

Berry, J.W., Poortinga, Y.H., Segall, M.H., & Dasen, P.R. (Eds.) (1992). *Cross-Cultural Psychology. Research and Applications*. Cambridge: University Press.

Cieciuch, J. & Davidov, E. (2016). Establishing measurement invariance across online and offline samples. A tutorial with the software packages Amos and Mplus. *Studia Psychologica* 15 (2), 83-99.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement Equivalence in Cross-National Research. *Annual Review of Sociology*, *40*, 55-75.

Davidov, E., Schmidt, P., Billiet, J. & Meuleman, B. (2018). *Cross-cultural analysis: Methods and applications.* New York: Routledge.

Davidov, E., & Siegers, P. (2010). Comparing basic human values in east and west Germany. In T. Beckers, K. Birkelbach, J. Hagenah & U. Rosar (Eds.), *Komparative empirische Sozialforschung [Comparative empirical Social Research]* (pp 43-63). Wiesbaden: VS Verlag.

Durkheim, E. (1964). *The Rules of the Sociological Method*. New York: The Free Press.

Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P. & Lyberg, L. et al. (2010). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts.* Hoboken (NJ): John Wiley & Sons.

Harkness, J.A., van de Vijver, F.J.R. & Mohler, P.P. (Eds.) (2003). *Cross-Cultural Survey Methods*. New York: John Wiley

Horn, J.L. & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Exp. Aging Res.* 18, 117-44.

Johnson ,T. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. In J.A. Harkness (Ed.), *Zuma-Nachrichten Spezial Volume 3. Cross-Cultural Survey Equivalence*, (pp. 1-40). Mannheim: Zuma.

Jowell, R., Roberts, C., Fitzgerald, R. & Eva, G. (2007). *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*. London: Sage.

Jowell, R. (1998). How comparative is comparative research? *Am. Behav. Sci.* 42(2), 168-77.

Lynn, P., Japec, L. & Lyberg, L. (2005). What's so special about cross-national surveys? In J.A. Harkness (Ed.) *Zuma-Nachrichten Spezial* Volume 12. *Conducting Cross-National and Cross-Cultural Surveys*, (pp. 7-20). Mannheim: Zuma.

Sarrasin, O., Green, E.G.T., Berchtold, A. & Davidov, E. (2012). Measurement equivalence across subnational groups: an analysis of the conception of nationhood in Switzerland. *International Journal of Public Opinion Research* 25(4): 522-534.

Widaman, K.F., Ferrer, E. & Conger R.D. (2010). Factorial invariance within longitudinal structural equation models: measuring the same construct across time. *Child Development Perspectives 4*(1): 10–18.

Acknowledgment

The second guest editor would like to thank the University of Zurich Research Priority Program "Social Networks" for their support during work on this special issue.

methods, data, analyses | Vol. 12(1), 2018, pp. 7-46 DOI: 10.12758/mda.2017.06

Testing the Measurement Invariance of Political Trust across the Globe. A Multiple Group Confirmatory Factor Analysis

Wiebke Breustedt
University of Duisburg-Essen/University of Cologne

Abstract

Today, comparative social scientists have ample survey data to test the generalizability of theories related to political trust. Unless its measurement invariance has been established, they run the risk of drawing invalid conclusions though. Based on different sets of items and dimensional models, previous studies have yielded diverging results regarding the measurement invariance of political trust in Europe and former Soviet countries. Using a set of six items and contrasting three competing dimensional models, this study tests the measurement invariance of political trust across the globe in 32 electoral and liberal democracies. It uses multiple group confirmatory factor analysis and draws on data from the World Values Survey (wave 6, 2010-2014). Configural invariance of a revised two-dimensional model of trust in implementing and representative political institutions was established in 19 democracies when excluding trust in civil service. Full invariance of this model was established in three post-communist countries in eastern and southeastern Europe. The results corroborate that the measurement invariance of political trust must not be assumed. Conceptually, they provide reason to infer that, by and large, people in democracies have a two-dimensional construct of political trust. Methodologically, they manifest that trust in civil service is an ambiguous item, which is not as meaningfully related to the construct of political trust as other items.

Keywords: measurement equivalence, measurement invariance, multiple group confirmatory factor analysis, political trust, trust in political institutions

Introduction

Today more than ever, comparative social scientists can test the generalizability of theories pertaining to the changes, sources, and consequences of political trust thanks to the growing availability of cross-national survey data (Braun, 2013; Zmerli & van der Meer, 2017). This is a decisive, but not a conclusive step forward. Unless the comparability of political trust measures has been established, inferences about the generalizability of political trust theories across the globe may be invalid (Davidov, Meulemann, Cieciuch, Schmidt, & Billiet, 2014).

The issue of comparability results from the fact that people's political trust is a construct. As such, it is a latent property of individuals that cannot be measured directly (Jackman, 2008). Cross-national researchers therefore have to rely on observed measures such as survey items pertaining to trust in different political objects. According to the 'response process model' (Tourangeau, Rips, & Rasinski, 2000), answers to these items allow inferences about people's underlying construct of political trust. Based on this assumption, studies commonly use political trust items to create additive or averaged index scores (see for example Catterberg & Moreno, 2006; Chang & Chu, 2006).

While indices are a common and convenient measurement instrument, the index scores are not necessarily comparable across countries and over time. A key to valid comparisons is to establish the invariance of the measurement instrument. "The general question of invariance of measurement is one of whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attributes" (Horn & Mcardle, 1992, p. 117). Various forms of bias may systematically distort the invariance of measures (van de Vijver & Tanzer, 2004). For example, asking about people's trust in a political institution such as civil service may be biased because civil service's responsibilities and tasks differ across countries. Or, owing to the translation of the response scales, the difference between 'a great deal of trust' as opposed to 'quite a lot of trust' may not be judged in the same way by respondents from different countries, thereby biasing their responses.

Because of these potential biases, it is essential to test the measurement invariance of the political trust items beforehand. The goal is to determine whether and

Acknowledgments

I would like to thank Eldad Davidov, Lluís Coromina, Cristiano Vezzoni, Susanne Pickel, Dominik Becker, Christina Zuber as well as the three anonymous reviewers for their helpful comments and suggestions.

Direct correspondence to

Wiebke Breustedt, University of Duisburg-Essen, Center for Higher Education and Quality Development, Keetmanstr. 3-9, 47058 Duisburg, Germany
E-mail: wiebke@breustedt.org

to what extent the proposed measurement model matches the observed structure of the data, thereby supporting the assumption that political trust can be measured across countries by a common set of items using the same number of latent factors (Milfont & Fischer, 2010). If measurement invariance is not tested beforehand, comparisons of observed differences in means may not reflect actual differences in people's average level of political trust and regression coefficients may suggest false relationships. In addition, true country-specific or temporal differences may be obscured (Chen, 2008). Either way, using political trust indices without testing for measurement invariance may lead to invalid conclusions regarding the changes, sources, and consequences of political trust (Ariely & Davidov, 2012; Vandenberg & Lance, 2000).

The lack of a common measurement model of political trust complicates such a test. First, there is no common set of political trust items and second, there is no agreement on the dimensionality of political trust.[1] This is best exemplified by previous cross-country exploratory studies (see Table 1). They reach different conclusions regarding the dimensionality of political trust depending on the estimation method and specifications, the design (pooled or country-specific), and the items used. This lack of consensus hampers valid comparisons.

Recently, several researchers tested the measurement invariance of political trust in European and former Soviet countries by means of multiple group confirmatory factor analysis. This method provides a stringent test because every element of the measurement model (not just the number of factors) is specified beforehand and the model outputs allow researchers to discern the reasons for invariance in detail (Brown, 2006). The studies tested and supported different dimensional models of political trust. Whereas some show that it is a single-dimensional construct, others provide evidence that a two-dimensional model of political trust in representative and implementing institutions reaches different levels of measurement invariance, depending on the countries of analysis and the chosen items (see Table 2).

Given these diverging measures and results, the question of the appropriate measurement model of political trust remains subject to debate. In addition, previous measurement invariance tests of political trust have focused on European and former Soviet countries, neglecting Asia, Africa, and Latin America. The purpose of this article is to determine: To what extent can the measurement invariance of political trust be established across the globe and if so, based on which measurement model?

1 The issue of comparability is further exacerbated by the fact that there is no uniform wording and response scale for political trust items.

Table 1 Previous Cross-Country Exploratory Analyses of the Dimensionality of Political Trust

author(s)	survey	time point/ period	countries	method	trust in							
					gov-ern-ment	parlia-ment	politi-cians	pol. parties	civil service	courts/ legal system	the police	the army
Hooghe & Kern (2015)	ESS	2002-2010	30 (Europe)	principal component analysis (pooled across countries and time)								
Rogge & Kittel (2014)	ESS	2008, 2010	29 (Europe)	principal component analysis (separately for each time point; pooled across countries)								
Lu (2014)	Asia Barometer Survey WVS	2006-2008 2005-2007	5 (Brazil, Russia, India, China, South Africa)	principal component analysis ((1) separately for each country but pooled over time; (2) pooled across countries and time)								
Braun (2013) (1)	EVS/WVS	1994-1999 2005-2007 2008	14 southern, eastern and southeastern European countries	principal component analysis (pooled by country groups; separately for each time point)								

Table 1 continued

author(s)	survey	time point/ period	countries	method	trust in							
					gov-ern-ment	parlia-ment	politi-cians	pol. parties	civil service	courts/ legal system	the police	the army
Zmerli (2013) (2)	ESS	2002, 2004, 2006	European countries (# varies by year)	principle component analysis (separately for each time point and country); specifies no. of components								
Marien (2011a)	ESS	2006, 2008	23 (Europe)	principal component analysis (pooled across countries and time)								
Newton & Zmerli (2011)	WVS	2005-2007	22 democracies	principal component analysis (pooled across countries)								
Oskarsson (2010)	ESS	2002, 2004	23 (Europe)	principal component analysis (pooled across countries)								
Rose & Mishler (2010)	NEB	1993-2004	14 post-Communist countries	principal component analysis (pooled across countries and time)								

Table 1 continued

author(s)	survey	time point/ period	countries	method	trust in							
					gov-ern-ment	parlia-ment	politi-cians	pol. parties	civil service	courts/ legal system	the police	the army
Slomczynski & Janicka (2009)	ESS	2006	23 (Europe)	factor analysis (pooled across countries and separately per country)								
Listhaug & Ringdal (2008)	ESS	2004	24 (Europe)	factor analysis (separately per country); specify no. of factors								
Zmerli & Newton (2008) (3)	ESS US CID	2002-2003 2006	23 (Europe), US	principal component analysis (separately for each country)								
Denters et al. (2007)	CID	1999-2002	13 (Europe)	factor analysis (pooled across countries); specify no. of factors								
Zmerli et al. (2007) (4)	CID	1999-2002	13 (Europe)	principal component analysis (separately for each country)								

Table 1 continued

author(s)	survey	time point/ period	countries	method	trust in							
					gov-ern-ment	parlia-ment	politi-cians	pol. parties	civil service	courts/ legal system	the police	the army
Lühiste (2006)	NBB	2001	3 (Latvia, Lithu-ania, Estonia)	principal com-ponent analysis (pooled across countries)								
Zmerli (2004) (3)	ESS	2002	21 (Europe)	principal com-ponent analysis (separately for each country)								
Fuchs et al. (2002) (5)	WVS	1995-1997	6 (Europe), US	factor analysis (pooled across countries); specify no. of factors								

Note. the analyses also include items measuring trust in (1) the press and unions; (2) European Parliament; (3) European Parliament and the UN; (4) municipal board; (5) 14 additional items measuring political support; the shading of the cells indicates the dimensional structure found in the analyses. Own compilation.

Table 2 Previous Multiple Group Confirmatory Factor Analyses of Political Trust

author(s)	survey	time point/ period	number of countries	level of measurement invariance reached	trust in										
					politicians (1)	political parties (2)	government (3)	parliament (4)	courts (5)	police (6)	regional government (7)	local government (8)	army (9)	EU parliament (10)	UN (11)
Coromina & Davidov (2013)	ESS	2002-2008	7 3	partial metric; partial scalar											
Marien (2011b)	ESS	2004 2006 2008	23 21 22	partial metric	ϕ_{21}	ϕ_{21}			ϕ_{65}	ϕ_{65}					
Marien (2017)	ESS	2012	23	partial scalar	ϕ_{21}	ϕ_{21}			ϕ_{65}	ϕ_{65}					
Schaap & Scheepers (2014)	ESS	2010	19	scalar*	ϕ_{21}	ϕ_{21}									
Ariely (2015)	EVS	2008	32	metric											
Schneider (2017)	LITS II	2010	23	metric											
Schneider (2017)	LITS II	2010	21	partial scalar											
Schneider (2017)	LITS II	2010	29	partial scalar						ϕ_{96}			ϕ_{96}		
Schneider (2017)	LITS II	2010	35	partial metric											
André (2014)	ESS	2008	22	partial scalar	ϕ_{21}	ϕ_{21}			ϕ_{65}	ϕ_{65}				ϕ_{1110}	ϕ_{1110}

Note. the shading of the cells shows the factor structure of political trust in the model tested by the author(s). ϕ indicates an error covariance between the respective items. For example, ϕ_{96} indicates an error covariance between trust in the army and trust in the police. * the analysis focused on the invariance of trust in the police. Own compilation.

The study extends previous analyses in several ways. First, it tests the measurement invariance of political trust on a global scale in 32 electoral and liberal democracies. Second, the analysis provides a detailed debate and conclusion regarding the dimensionality of the construct of political trust. Third, it discusses the suitability of the available items for cross-national comparisons in detail. Overall, the article's conclusions and recommendations can be used to inform future cross-national studies of political trust.

Since "any equivalence procedure can only be implemented successfully if an unambiguous specification of the concept is available" (van Deth, 2013, p. XXI), the article begins by defining political trust and by outlining three competing dimensional models of political trust. The subsequent section describes the research design and the three alternative measurement models of political trust that follow from the dimensional models. In the analysis section, the measurement invariance test of political trust is presented. The article concludes by outlining the implications of the findings and recommendations for the comparative study of political trust.

Competing Dimensional Models of Political Trust

Political trust can be defined as people's positive anticipatory expectation that, despite uncertainty, the conduct of the political trustee in question will be in line with their normative expectations (Miller & Listhaug, 1990; Möllering, 2006).[2] Researchers generally agree that trust in different political trustees such as parliament, the judiciary, and government can be distinguished theoretically (Levi & Stoker, 2000). They disagree on the empirical dimensionality of citizens' construct of political trust, though, resulting in three competing dimensional models.

The first dimensional model proposes a distinction between trust in political authorities and trust in political institutions. Building on Easton's (1975) classic model of political support, several researchers advocate that the two are related but separate dimensions of political trust (Dalton, 2004; Denters, Gabriel, & Torcal, 2007; Norris, 2011). First and foremost, they assume that people perceive abstract and specific trustees separately: Abstract political institutions are characterized by rules that define relationships among political roles, thereby prescribing and constraining the interactions of political actors in general over time; specific politi-

2 To date, there is no commonly accepted definition of political trust. Some conceptualize it as a kind of supportive behavior (Fisher, van Heerde, & Tucker, 2010) whereas others regard it as an attitude (Miller & Listhaug, 1990). Relatedly, the elements of the definitions of political trust that they stipulate do not coincide. Furthermore, some researchers state that the term 'trust' can 'travel' to political institutions without overstretching its conceptual core (Fuchs, Gabriel, & Völkl, 2002). Others maintain that 'trust' in political institutions should be referred to as 'confidence' (Hardin, 2000).

cal incumbents enact and interpret these roles within a particular period of time (March & Olsen, 1989). Second and consequently, while people may not trust the current political incumbents, they do not necessarily doubt that the conduct of the political institution in question will be in line with their normative expectations once the incumbents are no longer in office. At the same time, the two dimensions are related because incumbents affect the perception of the institutions. Proponents of this dimensional model assert that the distinction should be maintained all the same because it may yield more valid insights on the changes, sources, and consequences of political trust (Dalton, 2004; Norris, 2011).

According to the second dimensional model, the distinction between trust in representative and implementing political institutions is more plausible. Several researchers assume that citizens' political trust has two dimensions because people broadly categorize the responsibilities and characteristics of the work of political institutions into two groups. On the one hand, representative political institutions such as political parties, government, and parliament serve to make collectively binding decisions. By and large, their work is characterized by political controversies and competition. On the other hand, implementing political institutions such as the courts and police are responsible for maintaining order and implementing the law. On the whole, political partisanship is less prominent in their daily work (Gabriel, 1999; Pickel & Walz, 1995; Rothstein & Stolle, 2003). Within this group of researchers, there is disagreement regarding the attribution of trust in civil services, though. According to some, it is affected by people's overall trust in implementing political institutions as civil services serve to enact government policies (Gabriel, 1999). According to others, civil service officials may be perceived as agents of government precisely because they implement its laws, thereby politicizing the perception of the trustee (Rothstein & Stolle, 2008). This in turn may cause people to attribute it to their overall trust in representative political institutions. Leaving aside these differences, proponents of this two-dimensional model generally argue that trust in representative and implementing political institutions is related because the latter act on the basis of laws that were drafted and adopted by the former (Fuchs et al., 2002).

Still others have proposed a third, single-dimensional model of political trust. Some state that it especially applies to citizens in newly established democracies who have not had sufficient experience to distinguish between representative and implementing political institutions (Mishler & Rose, 1994). Others maintain that this model also holds in established democracies. This may be because individuals learn to trust at an early age and generalize this socialization experience to the political realm. People's generalized trust attitude is assumed to 'spill up' to political institutions (Mishler & Rose, 2001). Another line of argument suggests that political trust is "a comprehensive assessment of the political culture that is prevalent within a political system" (Hooghe 2011, p. 275). As a system characteristic,

political culture is assumed to impact political actors and institutions alike. As a result, people evaluate political objects and form political trust 'en bloc'. Therefore people are expected to trust political trustees to a greater or lesser extent without making more fine-grained distinctions.

These competing dimensional models suggest three alternative measurement models of political trust for the measurement invariance test. Depending on the dimensional model, the number of latent factors as well as the relational structure between the latent factors and observed items of political trust differ. These dimensional models were therefore translated into measurement models for the analysis.

Research Design

Operationalization

The analysis of the measurement invariance of political trust is based on data from the most recent wave of the World Values Survey (WVS). The WVS is the largest non-commercial, cross-national, time-series survey of public opinion and value preferences. Its most recent wave (wave 6, 2010-2014) covers 57 countries around the world and includes a number of items measuring trust in different political trustees, thereby permitting a measurement invariance test of political trust across the globe (World Values Survey, 2017). Since there is no common set of political trust items, the items that were used most frequently in previous studies of the dimensionality of political trust were selected from those available in the WVS (see Tables 1 and 2): trust in the police, the courts, the government, political parties, parliament, and civil service. The items are measured on an ordinal scale with four response categories. For each of the political trustees, WVS respondents were asked to indicate "how much confidence [they] have in that organization: a great deal of confidence, quite a lot of confidence, not very much confidence, or none at all". The same items were administered to the respondents in the respective national languages. This reduces the chance that the measurement invariance test reflects differences in item-wording rather than actual differences in respondents' construct of political trust across countries. The original data were recoded to include only one kind of missing value and to range from 0 (none at all) to 3 (a great deal of trust).

Case Selection

The study analyzed the measurement invariance of political trust in electoral and liberal democracies. Non-democratic states were excluded because citizens' relationship with and the functional interaction of political trustees such as government and the courts differ in these countries. These differences may impact the way

the construct of political trust develops in people's minds in democracies and non-democracies (Mishler & Rose, 1997).[3] This assumption is substantiated by Schneider's (2017) as well as Schaap and Scheepers' (2014) analysis of the measurement invariance of political trust in European and former Soviet countries. They found that a greater level of measurement invariance could be established once former Soviet autocracies were excluded from the analysis. The study at hand therefore focused on democracies in order to eliminate this possible source of measurement non-equivalence.

The countries included in the study were selected based on Polity IV (Center for Systemic Peace, 2016). Polity IV comprises indicators of institutional autocracy and democracy (Marshall, Gurr, & Jaggers, 2015). Countries' polity score can range from -10 (fully autocratic) to +10 (fully democratic). In line with the threshold provided on the Polity IV website (Marshall & Gurr, 2014), countries were included if their polity score was six or higher in the year the survey was conducted as well as four years prior to this year.

The final sample consisted of 32 countries with 46,315 respondents. The selected countries as well as the sample sizes and missings per item are listed in Table A1 in the appendix.[4] The survey samples are representative of the countries' adult population (World Values Survey, 2017).

3 As Breustedt and Stark (2015) argue, in authoritarian countries it is difficult for citizens to distinguish political institutions because of the lack of a system of checks and balances. In addition, as elections are infrequent or inconsequential, political institutions become mainly associated with the political incumbents. Therefore, people in authoritarian states most likely develop their trust in different political trustees in tandem. According to Rivetti and Cavatorta (2017), political trust in democratic regimes is positive whereas in authoritarian regimes it is negative: "whereas positive political trust can be defined as trust in ethical, legal or just actions undertaken by the ruling authority, negative trust can be defined as trust in the fact that the authority will act predictably" (Rivetti & Cavatorta, 2017, p. 60). Still, political trust in authoritarian countries is not necessarily devoid of positive normative expectations. People's normative expectations of political trustees may simply differ in authoritarian countries. Either way, measures of political trust in democracies and autocracies are not likely to be equivalent as responses to the same items are susceptible to construct bias.

4 Table A1 reports the original sample sizes. Most items have less than 5% missing per country. Two issues stand out: Trust in civil service has > 5% missing in nine countries, 18.4% of the cases for trust in government are missing in Lebanon, and Japan is the country with the largest amount of missing data. Cases were dropped if they had missings on all six items for the analysis. Respondents from the WVS wave 6 survey in India, conducted in 2012, were excluded because the wave 6 data file also includes a more recent Indian survey sample from 2014. 'Pairwise present' was used to handle missing data (Asparouhov & Muthén, 2010).

Method

The measurement invariance (MI) of political trust was tested using multiple group confirmatory factor analysis (MGCFA). Alternative methods include item response theory and latent class analysis (Davidov et al., 2014; Kankaraš, Vermunt, & Moors, 2011; Millsap, 2011). The study used MGCFA because it is a widely applied method to test MI and because previous studies of the MI of political trust used this method.

The analysis was conducted in three stages. Because there is no agreed upon measurement model of political trust, first, confirmatory factor analysis (CFA) was used to determine the model fit of the three alternative models derived from the dimensional models outlined above in each of the 32 countries. The best-fitting model served as the baseline model in the second step, the simultaneous analysis of MI across countries by means of MGCFA. Based on these empirical results as well as theoretical considerations, in the third step, this measurement model was revised and subsequently tested for MI.

Consonant with the three dimensional models described earlier, three measurement models were developed as possible baseline models for the MI test (see Figures 1 to 3).[5] Civil service was specified to load on trust in representative institutions in line with previous exploratory analyses (see Table 1). None of the models included any error correlations. In the two-dimensional models, the latent factors were assumed to correlate.

The study took account of the ordinal measurement scale of the items. Lubke and Muthén (2004) have shown that treating ordered-categorical data as continuous may yield estimates that suggest that the factor structure found in different countries differs when, in fact, it is equivalent. To circumvent this issue, the study followed a common approach to estimate latent variable models for ordered-categorical items – the latent response variable model (Muthén & Asparouhov, 2002).

This approach is outlined briefly as it affects the way MI tests are conducted. As indicated in Figures 1 to 3, the model estimation based on the latent response variable model assumes that the latent factor(s) of political trust (ξ_i) cause(s) the variance and covariance among latent response variables of political trust in six different political trustees (χ^*_i). The latent response variables are taken to have a continuous and normally distributed scale. Their relationship with the latent factor(s) is understood to be linear. Thus, as in standard MGCFA with continuous items, each latent response variable has a factor loading (λ_i), an intercept (τ_i), and an error term. The latent response variables are assumed to be the unobserved

5 Some researchers have distinguished between trust in political actors, representative political institutions, and implementing political institutions (Denters et al., 2007; Gabriel, 1999). This three-dimensional model could not be tested because of the limited number of survey items available in the WVS.

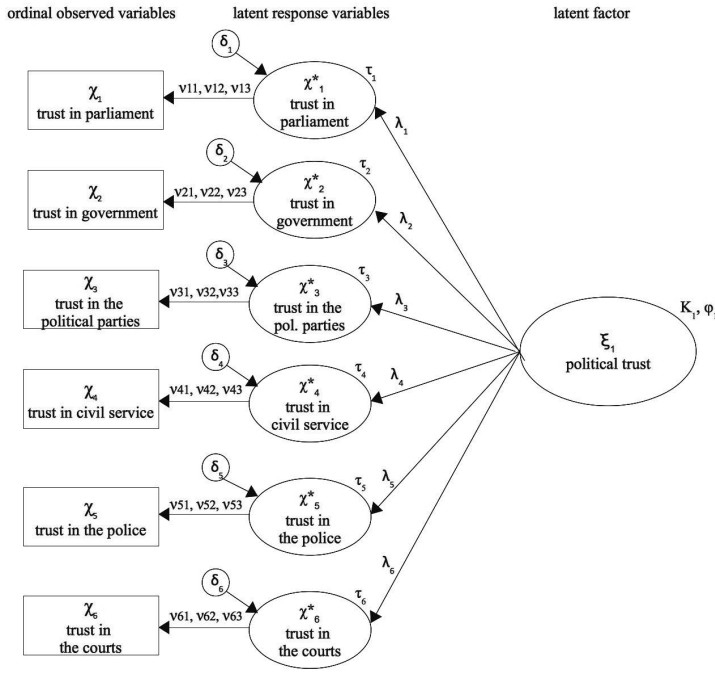

Figure 1 Single-dimensional measurement model of political trust. Adapted from Davidov et al. (2011) and Poznyak et al. (2014). ξ (ksi): latent factor, κ (kappa): latent mean, φ (phi): factor variance, λ (lambda): factor loading, χ* (chi): latent response variable, τ (tau): intercept, δ (delta): error variance, χ (chi): observed variable, ν (nu): threshold.

latent counterparts of the observed ordered-categorical items of political trust (χ_i). The continuous nature of the latent response variables is roughly captured by the ordered-categorical response scale of the respective observed items. Each pair of response categories of the items represents a section of the continuous scale of the corresponding latent response variable. Each section therefore ends with a threshold (ν_{ij}). As a result, each latent response variable is related to its corresponding observed item through a set of thresholds, whereby the number of thresholds corresponds to the number of response categories minus one. Since the political trust items have four ordered response categories, the latent response variables each have three thresholds. That is to say, if χ_1 represents the ordinal item of trust in parliament and χ^*_1 stands for the latent response variable of trust in parliament, χ^*_1 reflects the amount of political trust needed to select a certain response category of χ_1. An observed response of '0' (none at all) in trust in parliament is expected if the level of χ^*_1 is less than or equal to the first threshold ν_{11}. If χ^*_1 is greater than ν_{11} but less than or equal to the second threshold ν_{12}, the predicted response is '1' (not very

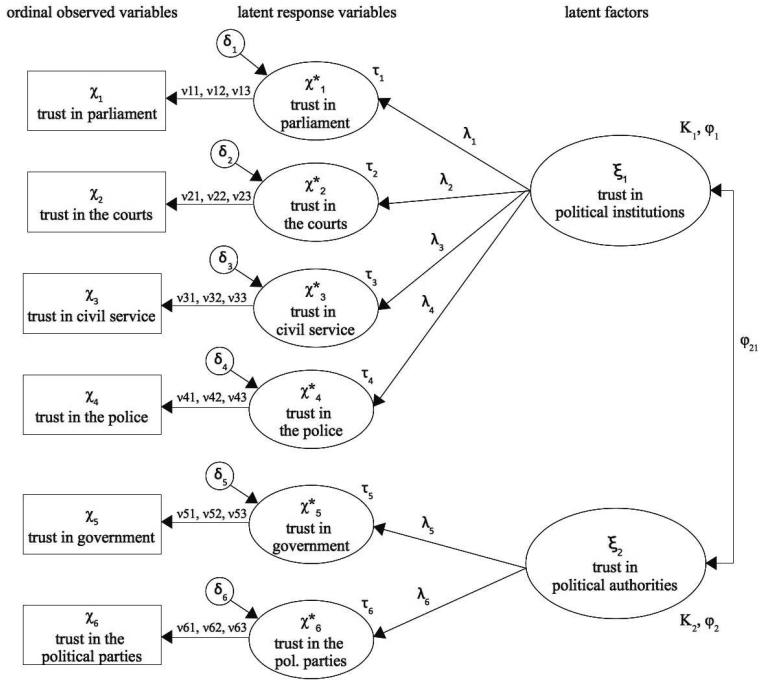

Figure 2 Two-dimensional measurement model of trust in political authorities and political institutions. Adapted from Davidov et al. (2011) and Poznyak et al. (2014). ξ (ksi): latent factor, κ (kappa): latent mean, φ (phi): factor variance, λ (lambda): factor loading, χ* (chi): latent response variable, τ (tau): intercept, δ (delta): error variance, χ (chi): observed variable, ν (nu): threshold.

much confidence). If the latent response variable of trust in parliament χ^*_1 is greater than ν_{12} but less than or equal to the third threshold ν_{13}, the predicted response is '2' (quite a lot of confidence). $\chi^*_1 > \nu_{13}$ corresponds to a response of '3' (a great deal of confidence) (Byrne, 2012; Kline, 2016; Millsap & Yun-Tein, 2004; Muthén & Asparouhov, 2002).

Accounting for the ordinal nature of the political trust items affects the parameters that have to be invariant across countries in order for MI to hold and, relatedly, the levels of MI that can be tested. The invariance of factor loadings, intercepts, and (unlike in the case of continuous variables) thresholds has to be considered (Davidov, Datler, Schmidt, & Schwartz, 2011; Millsap & Yun-Tein, 2004). Researchers can test to what extent these parameters are invariant by applying increasingly restrictive equality constraints in MGCFA and examining the respective model fit by means of goodness-of-fit indices. In the case of ordered-categorical data, only two levels of MI are tested, namely configural and full MI (Davidov et al., 2011).

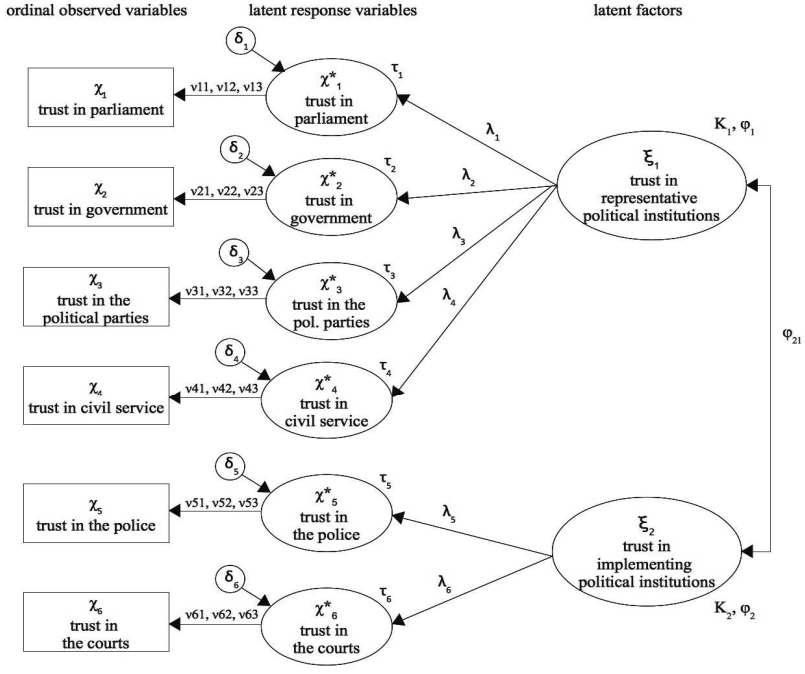

Figure 3 Two-dimensional measurement model of trust in representative and implementing political institutions. Adapted from Davidov et al. (2011) and Poznyak et al. (2014). ξ (ksi): latent factor, κ (kappa): latent mean, φ (phi): factor variance, λ (lambda): factor loading, χ* (chi): latent response variable, τ (tau): intercept, δ (delta): error variance, χ (chi): observed variable, ν (nu): threshold.

When testing for configural invariance, the estimated parameters are allowed to differ across countries. The test shows whether the number of factors and the pattern of fixed and free item factor loadings is the same across countries (Vandenberg & Lance, 2000). If this model fits the data, it may be inferred that people in different countries respond to political trust items with the same construct in mind (Chen, 2008). If not, country-specific measures may be required (Pendergast, von der Embse, Kilgus, & Eklund, 2017). Configural invariance is a prerequisite for full MI. Full MI requires the unstandardized factor loadings, intercepts, and thresholds to be equal (Davidov et al., 2011). If full MI is supported by the data, it can be inferred that the items measure the same latent construct, albeit with different degrees of precision because the error variances and covariances were not constrained to be equal (Kline, 2016). In addition, full MI implies that people in the

respective countries use the response scale in the same manner (Poznyak, Meulemann, Abts, & Bishop, 2014).[6]

The ordered-categorical nature of the data has a bearing on the appropriate choice of the method of estimation. As Brown (2006) notes, ignoring the fact that the data may be non-normally distributed could lead to incorrect parameter estimates, standard errors, and test statistics. The analyses were therefore run with the mean- and variance-adjusted weighted least squares (WLSMV) estimator in Mplus (Version 8) using the raw data. This estimator provides robust standard errors and (more) accurate estimates of factor loadings as well as corrected model test statistics. As Beauducel and Herzberg (2006) showed, it is superior to maximum likelihood estimation especially when the number of response categories is small, as in the case of the present study.

In order to conduct MI analyses, the scale of the latent factors has to be defined. Because latent factors are unobserved, they have no definite metric scale. In MGCFA, there are two common ways to establish this scale – the reference indicator method and the fixed factor method. When using the latter, the factor variances of the latent factors are fixed to one in all countries. This assumes that the factor variances are equal across countries. When applying the former, one factor loading per latent factor is fixed to one in all countries. Here the assumption is that this factor loading is invariant (Byrne, 2012). With regard to political trust, there is no evidence to justify either assumption. In this study, the reference indicator method was used because it was more straightforward to make a case for using single reference indicators.[7]

---

6 Unlike in the case of continuous data, the invariance of factor loadings alone does not establish comparability of the political trust measure because the item probability curves depend on the factor loadings, intercepts, and thresholds (Davidov et al., 2011; Muthén & Asparouhov, 2002). As a result, only two levels of measurement invariance were tested unlike in previous measurement invariance tests of political trust (Table 2). See Bowen and Masa (2015) for a summary of arguments in favor and against this practice.

7 In order to choose appropriate reference indicators, two exploratory factor analyses (EFA) were carried out per country (principal axis extraction; promax rotation). In the single-factor EFA, trust in parliament was the marker item in 22 out of 32 countries. In the two-factor EFA, in 28 out of 32 countries, trust in parliament was the item that loaded most strongly on one latent factor and in 17 out of 32 countries, trust in the police was the marker item of the other latent factor. Consequently, trust in parliament was used as the reference indicator in the single-dimensional model and trust in parliament as well as trust in the police were used as reference indicators in the two-dimensional model of trust in implementing and representative institutions. Trust in parliament and trust in government were used as reference indicators in the two-dimensional model of trust in political authorities and institutions. Trust in government was chosen because the author deemed it more likely that government is perceived in a comparable manner across countries compared to political parties because its structure and functions are more similar, differences notwithstanding. Table A2 in the appendix includes a robust-

Depending on the level of MI tested, additional parameters have to be fixed in order for the measurement model to be identified. The choice depends in part on the computer program and the model parameterization. Mplus was chosen because of its flexibility when testing the invariance of ordered-categorical items (Millsap & Yun-Tein, 2004). In practice, thresholds (ν_i) and intercepts (τ_i) cannot be estimated simultaneously. By default, Mplus fixes all intercepts of the latent response variables to zero, thereby allowing researchers to test the MI of thresholds (Davidov et al., 2011). In addition, Mplus offers two parameterization methods – delta and theta parameterization. Unlike delta parameterization, theta parameterization includes error variances for the latent response variables (δ) as estimated parameters (Muthén & Muthén, 1998-2017). This study used theta parameterization as previous MGCFAs (see Table 2) indicated that the error variances of some of the items might be correlated. In order to identify the measurement models, the following parameters were fixed. In the configural invariance model, one factor loading per latent factor as well as the error variances were fixed to one and the factor means were fixed to zero in all countries. In the full MI model, one factor loading per latent factor was fixed to one in all countries and the remaining factor loadings as well as the thresholds were constrained to be equal. In addition, the error variances were fixed to one and the factor means were fixed to zero in the reference country[8] and freely estimated in the other countries (Muthén & Muthén, 1998-2017).

The overall fit of the measurement models to the data was evaluated according to several criteria. X^2 as the classic fit index indicates exact fit between the estimated model parameters and the observed data. While this is informative, it is an unduly strong assumption for real-world data. In addition, X^2 is sensitive to sample size (Byrne, 2012; Meade, Johnson, & Braddy, 2008). Consequently, the goodness of fit evaluation was informed by the X^2 results but focused on three additional fit indices: the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis-Index (TLI). The 90% confidence interval of the RMSEA is provided to show how precise its point estimates are (MacCallum, Browne, & Sugawara, 1996). Following Yu (2002), the following cut-off criteria were used: TLI $\geq$0.95, CFI $\geq$0.96, and RMSEA $\leq$ 0.05.

The analysis also considered focal areas of ill fit. The proportion of variance of the indicator explained by the latent factor ('R-Square' in Mplus) was used to evaluate whether the items were meaningfully related to the respective latent factor. The extent of the correlation between the latent factors was taken into account to determine discriminant validity between the latent factors in case of the two-dimensional models of political trust (Brown, 2006). In addition, the study followed a dual modal two-pronged strategy proposed by Byrne and van de Vijver (2010).

ness test for Model A of the MGCFA (see Table 7). The analysis was not sensitive to the selection of these reference indicators.

8 Model C2: Australia; Model C3: Poland.

They suggest looking for patterns of misspecification that indicate that individual items, individual countries or groups of countries are the reason for measurement non-invariance. Modification indices, which approximate how much the model fit (X^2) would improve if the constrained or fixed parameter in question was freely estimated, can be used to discern such patterns (Brown, 2006). Because of X^2's sensitivity to sample size, it was considered in tandem with the respective expected parameter of change (EPC) value. Overall, these criteria provided information on the fit of the measurement models as well as how to revise the measurement models in order to establish full invariance.

Analysis

Establishing the Baseline Model of Political Trust

The first step in testing the MI of political trust on a global scale was to establish the baseline model. Tables 3 to 5 present the overall goodness-of-fit indices for each of the three alternative measurement models tested separately in 32 countries. In terms of CFI and TLI, the two-factor model of trust in political authorities and political institutions yielded the worst fit. As shown in Table 3, the two indices were above the recommended cut-off value in only five out of 32 countries. The RMSEA did not support the model in any of the countries. The latent covariance matrix of the factors was not positive definite in six countries. In all six countries, this was because the latent factor correlation was estimated to have an out of range value (> 1.0), signifying model misspecification because some or all of the items of one latent factor were more strongly related to some or all of the items of the other latent factor (Brown, 2006). In comparison, the single-factor model of political trust fit the data better (see Table 4). The CFI and TLI indicated good model fit in eight out of 32 countries. Finally, the two-factor model of trust in implementing and representative political institutions fit the data best (see Table 5). In 28 out of 32 countries, the CFI and TLI were above the recommended cut-off values. Furthermore, only in this model was the RMSEA smaller than 0.05 in two countries and its confidence interval indicated a good precision of this point estimate.

Table 3 Fit Measures for the Two-Factor Confirmatory Factor Analysis of Trust in Political Authorities and Political Institutions

country	n	χ^2 (df)	p-value	CFI	TLI	RMSEA (90% CI)	summary
all countries	46315	17403.165 (8)	0.00	0.953	0.912	0.217 (0.214-0.219)	
Argentina	1025	330.017 (8)	0.00	0.956	0.917	0.198 (0.180-0.217)	
Australia	1453	336.644 (8)	0.00	**0.966**	0.936	0.168 (0.153-0.184)	
Brazil	1486	the latent variable covariance matrix is not positive definite					
Chile	999	the latent variable covariance matrix is not positive definite					
Colombia	1509	the latent variable covariance matrix is not positive definite					
Cyprus	999	437.876 (8)	0.00	0.941	0.890	0.232 (0.214-0.251)	
Estonia	1531	781.502 (8)	0.00	0.948	0.902	0.251 (0.237-0.266)	
Georgia	1185	759.328 (8)	0.00	**0.965**	0.935	0.282 (0.265-0.299)	
Germany	2043	715.828 (8)	0.00	**0.960**	0.925	0.208 (0.195-0.221)	
Ghana	1552	the latent variable covariance matrix is not positive definite					
India	1578	149.767 (8)	0.00	0.880	0.774	0.106 (0.092-0.121)	
Japan	2350	1467.502 (8)	0.00	**0.975**	**0.954**	0.279 (0.267-0.291)	(√)
Lebanon	1183	68.742 (8)	0.00	**0.979**	**0.961**	0.080 (0.063-0.098)	(√)
Malaysia	1299	the latent variable covariance matrix is not positive definite					
Mexico	2000	410.193 (8)	0.00	**0.972**	0.947	0.159 (0.146-0.172)	
Netherlands	1849	818.027 (8)	0.00	**0.982**	**0.967**	0.234 (0.221-0.248)	(√)
New Zealand	812	236.709 (8)	0.00	**0.962**	0.930	0.188 (0.167-0.209)	
Peru	1206	291.760 (8)	0.00	**0.971**	0.945	0.171 (0.155-0.189)	
Philippines	1200	438.337 (8)	0.00	0.940	0.888	0.212 (0.195-0.229)	
Poland	957	304.620 (8)	0.00	**0.968**	0.939	0.197 (0.178-0.216)	
Romania	1488	742.378 (8)	0.00	**0.960**	0.924	0.248 (0.233-0.264)	
Slovenia	1060	298.563 (8)	0.00	**0.980**	**0.963**	0.185 (0.167-0.203)	(√)
South Africa	3477	973.607 (8)	0.00	**0.971**	0.946	0.186 (0.177-0.196)	
South Korea	1198	the latent variable covariance matrix is not positive definite					
Spain	1180	287.923 (8)	0.00	0.943	0.894	0.172 (0.155-0.190)	
Sweden	1205	516.348 (8)	0.00	0.948	0.902	0.230 (0.213-0.247)	
Taiwan	1204	224.002 (8)	0.00	**0.976**	**0.956**	0.150 (0.133-0.167)	(√)
Trinidad and Tobago	994	503.494 (8)	0.00	**0.960**	0.926	0.250 (0.231-0.268)	
Turkey	1593	528.707 (8)	0.00	0.951	0.909	0.202 (0.188-0.217)	
Ukraine	1500	934.882 (8)	0.00	**0.968**	0.941	0.278 (0.263-0.293)	
United States	2205	1429.113 (8)	0.00	0.931	0.871	0.284 (0.272-0.296)	
Uruguay	995	431.481 (8)	0.00	0.943	0.893	0.231 (0.212-0.249)	

Note. WLSMV estimator (theta parameterization), pairwise present was used to handle missing data (Asparouhov & Muthén, 2010), df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis-Index, RMSEA = root mean square error of approximation, 90% CI = 90% confidence interval, parameter of fit values above the recommended thresholds (Yu, 2002) are in bold, summary (√) indicates that two out of three fit indices are above the recommended thresholds, summary √ indicates that CFI, TLI, and RMSEA are above the recommended thresholds. Data are from the World Values Survey 2010-2014, 32 countries.

Table 4 Fit Measures for the Single-Factor Confirmatory Factor Analysis of Political Trust

country	n	χ^2 (df)	p-value	CFI	TLI	RMSEA (90% CI)	summary
all countries	46315	18131.958 (9)	0.00	0.951	0.919	0.209 (0.206-0.211)	
Argentina	1025	339.428 (9)	0.00	0.954	0.924	0.189 (0.172-0.207)	
Australia	1453	342.404 (9)	0.00	**0.965**	0.942	0.160 (0.145-0.174)	
Brazil	1486	467.487 (9)	0.00	0.947	0.911	0.185 (0.171-0.200)	
Chile	999	194.345 (9)	0.00	**0.977**	**0.962**	0.144 (0.126-0.161)	(√)
Colombia	1509	603.427 (9)	0.00	0.951	0.919	0.209 (0.195-0.224)	
Cyprus	999	478.871 (9)	0.00	0.936	0.893	0.229 (0.211-0.246)	
Estonia	1531	803.514 (9)	0.00	0.946	0.911	0.240 (0.226-0.254)	
Georgia	1185	804.307 (9)	0.00	**0.963**	0.938	0.273 (0.257-0.289)	
Germany	2043	739.886 (9)	0.00	0.959	0.931	0.199 (0.187-0.212)	
Ghana	1552	519.222 (9)	0.00	0.931	0.885	0.191 (0.177-0.205)	
India	1578	158.753 (9)	0.00	0.873	0.788	0.103 (0.089-0.117)	
Japan	2350	1593.134 (9)	0.00	**0.973**	**0.956**	0.274 (0.262-0.285)	(√)
Lebanon	1183	81.557 (9)	0.00	**0.975**	**0.959**	0.083 (0.067-0.099)	(√)
Malaysia	1299	878.559 (9)	0.00	0.955	0.925	0.273 (0.258-0.288)	
Mexico	2000	411.296 (9)	0.00	**0.972**	**0.953**	0.149 (0.137-0.162)	(√)
Netherlands	1849	891.088 (9)	0.00	**0.981**	**0.968**	0.230 (0.218-0.243)	(√)
New Zealand	812	245.580 (9)	0.00	**0.961**	0.935	0.180 (0.161-0.200)	
Peru	1206	294.694 (9)	0.00	**0.971**	**0.951**	0.162 (0.147-0.178)	(√)
Philippines	1200	437.427 (9)	0.00	0.940	0.901	0.199 (0.183-0.215)	
Poland	957	319.692 (9)	0.00	**0.966**	0.944	0.190 (0.172-0.208)	
Romania	1488	768.958 (9)	0.00	0.958	0.930	0.238 (0.224-0.253)	
Slovenia	1060	339.944 (9)	0.00	**0.978**	**0.963**	0.186 (0.170-0.203)	(√)
South Africa	3477	1041.826 (9)	0.00	**0.969**	0.949	0.182 (0.172-0.191)	
South Korea	1198	814.982 (9)	0.00	**0.964**	0.940	0.273 (0.258-0.289)	
Spain	1180	395.232 (9)	0.00	0.922	0.870	0.191 (0.175-0.207)	
Sweden	1205	546.657 (9)	0.00	0.945	0.908	0.223 (0.207-0.239)	
Taiwan	1204	222.983 (9)	0.00	**0.977**	**0.961**	0.141 (0.125-0.157)	(√)
Trinidad and Tobago	994	546.575 (9)	0.00	0.957	0.928	0.245 (0.228-0.263)	
Turkey	1593	570.242 (9)	0.00	0.948	0.913	0.198 (0.184-0.212)	
Ukraine	1500	1003.718 (9)	0.00	**0.966**	0.943	0.271 (0.257-0.286)	
United States	2205	1479.265 (9)	0.00	0.929	0.882	0.272 (0.261-0.284)	
Uruguay	995	442.719 (9)	0.00	0.942	0.903	0.220 (0.203-0.238)	

Note. WLSMV estimator (theta parameterization), pairwise present was used to handle missing data (Asparouhov & Muthén, 2010), df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis-Index, RMSEA = root mean square error of approximation, 90% CI = 90% confidence interval, parameter of fit values above the recommended thresholds (Yu, 2002) are in bold, summary (√) indicates that two out of three fit indices are above the recommended thresholds, summary √ indicates that CFI, TLI, and RMSEA are above the recommended thresholds. Data are from the World Values Survey 2010-2014, 32 countries.

Table 5 Fit Measures for the Two-Factor Confirmatory Factor Analysis of Political Trust in Implementing and Representative Political Institutions

country	n	χ^2 (df)	p-value	CFI	TLI	RMSEA (90% CI)	summary
all countries	46315	4004.959 (8)	0.000	**0.989**	**0.980**	0.104 (0.101-0.107)	(√)
Argentina	1025	25.885 (8)	0.001	**0.998**	**0.995**	**0.047** (0.027-0.067)	√
Australia	1453	149.490 (8)	0.00	**0.985**	**0.972**	0.110 (0.095-0.126)	(√)
Brazil	1486	278.099 (8)	0.00	**0.969**	0.941	0.151 (0.136-0.166)	
Chile	999	195.118 (8)	0.00	**0.977**	**0.956**	0.153 (0.135-0.172)	(√)
Colombia	1509	522.132 (8)	0.00	0.958	0.921	0.206 (0.192-0.222)	
Cyprus	999	82.736 (8)	0.00	**0.990**	**0.981**	0.097 (0.078-0.116)	(√)
Estonia	1531	221.914 (8)	0.00	**0.986**	**0.973**	0.132 (0.117-0.147)	(√)
Georgia	1185	316.563 (8)	0.00	**0.986**	**0.973**	0.180 (0.164-0.198)	(√)
Germany	2043	128.285 (8)	0.00	**0.993**	**0.987**	0.086 (0.073-0.099)	(√)
Ghana	1552	168.182 (8)	0.00	**0.978**	**0.960**	0.114 (0.099-0.129)	(√)
India	1578	129.277 (8)	0.00	0.897	0.807	0.098 (0.084-0.113)	
Japan	2350	117.045 (8)	0.00	**0.998**	**0.997**	0.076 (0.064-0.089)	(√)
Lebanon	1183	28.580 (8)	0.00	**0.993**	**0.987**	**0.047** (0.029-0.066)	√
Malaysia	1299	556.899 (8)	0.00	**0.972**	0.947	0.230 (0.214-0.246)	
Mexico	2000	211.765 (8)	0.00	**0.986**	**0.973**	0.113 (0.100-0.126)	(√)
Netherlands	1849	213.724 (8)	0.00	**0.995**	**0.992**	0.118 (0.105-0.132)	(√)
New Zealand	812	48.940 (8)	0.00	**0.993**	**0.987**	0.079 (0.059-0.101)	(√)
Peru	1206	102.030 (8)	0.00	**0.990**	**0.982**	0.099 (0.082-0.116)	(√)
Philippines	1200	187.409 (8)	0.00	**0.975**	**0.953**	0.137 (0.120-0.154)	(√)
Poland	957	96.655 (8)	0.00	**0.990**	**0.982**	0.108 (0.089-0.127)	(√)
Romania	1488	195.538 (8)	0.00	**0.990**	**0.981**	0.126 (0.111-0.141)	(√)
Slovenia	1060	56.482 (8)	0.00	**0.997**	**0.994**	0.076 (0.058-0.095)	(√)
South Africa	3477	467.079 (8)	0.00	**0.986**	**0.975**	0.128 (0.119-0.139)	(√)
South Korea	1198	564.953 (8)	0.00	**0.975**	**0.953**	0.241 (0.224-0.258)	(√)
Spain	1180	156.665 (8)	0.00	**0.970**	0.944	0.125 (0.109-0.143)	
Sweden	1205	98.056 (8)	0.00	**0.991**	**0.983**	0.097 (0.080-0.114)	(√)
Taiwan	1204	112.167 (8)	0.00	**0.989**	**0.979**	0.104 (0.087-0.121)	(√)
Trinidad and Tobago	994	102.419 (8)	0.00	**0.992**	**0.986**	0.109 (0.091-0.128)	(√)
Turkey	1593	204.398 (8)	0.00	**0.982**	**0.966**	0.124 (0.110-0.139)	(√)
Ukraine	1500	108.100 (8)	0.00	**0.997**	**0.994**	0.091 (0.076-0.107)	(√)
United States	2205	537.652 (8)	0.00	**0.974**	**0.952**	0.173 (0.161-0.186)	(√)
Uruguay	995	54.921 (8)	0.00	**0.994**	**0.988**	0.077 (0.058-0.097)	(√)

Note. WLSMV estimator (theta parameterization), pairwise present was used to handle missing data (Asparouhov & Muthén, 2010), df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis-Index, RMSEA = root mean square error of approximation, 90% CI = 90% confidence interval, parameter of fit values above the recommended thresholds (Yu, 2002) are in bold, summary (√) indicates that two out of three fit indices are above the recommended thresholds, summary √ indicates that CFI, TLI, and RMSEA are above the recommended thresholds. Data are from the World Values Survey 2010-2014, 32 countries.

At the same time, the inspection of focal areas of ill fit of the CFAs of the two-factor model of trust in implementing and representative political institutions suggested several items and countries of concern. X^2 strongly varied across countries, ranging from 564.953 in South Korea to 25.885 in Argentina (see Table 5). The standardized correlation coefficient between the two latent factors was > .85 in five countries, indicating low discriminant validity (see Table 6). These aspects point to possible countries as a reason for measurement non-invariance. As for the items, 'trust in civil service' was the item with the lowest proportion of explained variance in 21 countries (see Table 6). In addition, the modification and expected parameter change indices recommended a positive cross-loading between the latent factor 'trust in implementing political institutions' and the item 'trust in civil service' in 17 countries. In 13 countries, this modification index value was the largest among all suggested cross-loadings between a latent factor of political trust and a political trust item (see Table 6). This indicates that 'trust in civil service' is an ambiguous item not as meaningfully related to the construct of political trust as the other items. Furthermore, in 22 countries, the modification and expected parameter change indices for error co-variances pointed out that the model fit would improve if a cross-loading were added between 'trust in parliament' and 'trust in political parties'. This modification index was the largest value for suggested error correlations in nine countries (see Table 6).

Based on these results, the two-factor model of trust in implementing and representative political institutions was chosen as the baseline model for the MGCFA. The focal areas of ill fit informed its revision for the MI test across countries.

Table 6 Focal Areas of Ill Fit in the Two-Factor Confirmatory Factor Analysis of Political Trust in Implementing and Representative Political Institutions per Country

	focal areas of ill fit			
country	political trust item with the lowest explained variance	largest modification index for cross-loadings between a latent factor of political trust and a political trust item*	largest modification index for error correlation*	factor correlation >.85
Argentina	police	---	parliament and police (neg.)	
Australia	police	ξ2 **and civil service**	**parliament and political parties**	
Brazil	**civil service**	ξ2 and government	**parliament and political parties**	
Chile	police	ξ2 and government	government and police	x
Colombia	police	ξ2 and government	government and court	x
Cyprus	**civil service**	ξ2 and government	civil service and parliament	
Estonia	**civil service**	ξ2 **and civil service**	**parliament and political parties**	
Georgia	**civil service**	ξ2 **and civil service**	civil service and police	
Germany	**civil service**	ξ2 **and civil service**	civil service and police	
Ghana	**civil service** and political parties	ξ2 and government	government and court	
India	political parties	ξ2 and government	**parliament and political parties**	
Japan	court	ξ2 **and civil service**	political parties and government	
Lebanon	**civil service**	---	parliament and government (neg.)	
Malaysia	**civil service**	ξ2 and government	**parliament and political parties**	x
Mexico	police	ξ2 and government	government and court	
Netherlands	**civil service**	ξ2 **and civil service**	civil service and police	
New Zealand	**civil service**	ξ2 and political parties (neg.)	political parties and court (neg.)	

Table 6 continued

country	focal areas of ill fit			
	political trust item with the lowest explained variance	largest modification index for cross-loadings between a latent factor of political trust and a political trust item*	largest modification index for error correlation*	factor correlation >.85
Peru	police	ξ_2 and government	government and court	
Philippines	**civil service**	ξ_2 and government	government and court	
Poland	police	ξ_2 **and civil service**	**parliament and political parties**	
Romania	**civil service**	ξ_2 **and civil service**	civil service and police	
Slovenia	police	---	civil service and parliament	
South Africa	**civil service**	ξ_2 and government	government and court	x
South Korea	**civil service**	ξ_2 and government	**parliament and political parties**	
Spain	**civil service**	ξ_2 **and civil service**	political parties and government	
Sweden	**civil service**	ξ_2 **and civil service**	civil service and court	
Taiwan	**civil service**	ξ_2 and government	**parliament and political parties**	x
Trinidad and Tobago	**civil service**	ξ_2 **and civil service**	civil service and court	
Turkey	political parties	ξ_2 **and civil service**	civil service and government (neg.)	
Ukraine	**civil service**	ξ_2 and government	**parliament and political parties**	
United States	**civil service**	ξ_2 **and civil service**	civil service and police	
Uruguay	**civil service**	ξ_2 and government	government and court	

Note. ξ_2 = latent factor of trust in implementing political institutions, * positive expected parameter change unless otherwise indicated, (neg.) = expected parameter change is negative

Table 7 Fit Measures for the Multiple Group Confirmatory Factor Analysis of Political Trust

Model	χ^2 (df)	p-value	CFI	TLI	RMSEA (90% CI)
Model A *(all items and countries)*					
1. Configural invariance	6457.907 (256)	0.00	**0.987**	**0.976**	0.129 (0.127-0.132)
Model B *(excluding trust in civil service)*					
1. Configural invariance	3915.855 (128)	0.00	**0.991**	**0.978**	0.143 (0.139-0.147)
Model C1 *(excluding trust in civil service, correlated error between trust in parliament and trust in political parties)*					
1. Configural invariance	919.890 (96)	0.00	**0.998**	**0.994**	0.077 (0.073-0.082)
Model C2 *(excluding trust in civil service, correlated errors between trust in parliament and trust in political parties, including Australia, Brazil, Cyprus, Estonia, Georgia, Germany, Ghana, India, Japan, New Zealand, Philippines, Poland, Romania, Slovenia, South Korea, Sweden, Trinidad & Tobago, Ukraine, Uruguay)*					
1. Configural invariance	235.782 (57)	0.00	**0.999**	**0.998**	**0.048 (0.042-0.055)**
2. Full invariance	5430.023 (255)	0.00	**0.980**	**0.985**	0.123 (0.120-0.126)
Model C3 *(excluding trust in civil service, correlated errors between trust in parliament and trust in political parties, including Poland, Romania, Slovenia)*					
2. Full invariance	115.991 (31)	0.00	**0.998**	**0.998**	**0.048 (0.039-0.058)**

Note. WLSMV estimator (theta parameterization), pairwise present was used to handle missing data (Asparouhov & Muthén, 2010), df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis-Index, RMSEA = root mean square error of approximation, 90% CI = 90% confidence interval, parameter of fit values above the recommended thresholds (Yu, 2002) are in bold. Data are from the World Values Survey 2010-2012, 32 countries.

Testing the Measurement Invariance of Political Trust

Table 7 presents the results of the MI test of political trust in 32 democracies across the globe. Initially, the configural invariance of the baseline model was tested (Model A). While the CFI and TLI indicated good model fit, the RMSEA was well above the cut-off criterion. Paying heed to the focal areas of ill fit that were discerned in the single-country CFAs (see Tables 5 and 6), trust in civil service was

excluded from the measurement model (Model B). This improved the CFI and TLI somewhat and the X^2 notably.

Again based on the findings from the single-country CFAs, errors of trust in parliament and trust in political parties were then allowed to correlate (Model C1). This error correlation indicates that the two measurement errors are systematically related because some of the shared variance of the two items is due to another common outside cause. Substantively, most likely, this is because political parties play a major role in parliament unlike in the other political institutions. The model adjustment considerably improved the X^2, the CFI and TLI as well as the RMSEA. The latter remained above the recommended cutoff criterion, however.

Based on the results of Model C1, 13 countries were excluded because of model fit issues – eight countries because the factor correlation exceeded .85[9], two countries because the cross-loading between trust in parliament and trust in political parties was not significant (Argentina)[10] or negative (Spain) and three countries because the highest modification index indicated ill specification owing to a missing cross-loading between the latent factor trust in implementing institutions and trust in political parties (Netherlands: 158.388, Turkey: 69.156), and trust in government and trust in the courts (USA: 161.571) (Model C2). Model C2 – including 19 electoral and liberal democracies – reached configural invariance. In all of these countries, the model fit the data well: the unstandardized factor loadings and error correlation were significant at the .05 level; the size of the completely standardized factor loadings was substantial and their direction positive, as expected; the completely standardized factor correlations were all <.85; the error variances were positive and the modification indices were all < 26. Model C2 did not reach full invariance, however.[11]

When the data do not support full invariance, researchers have several options (Davidov, Dülmer, Schlüter, Schmidt, & Meulemann, 2012). A popular strategy is to test for partial MI, that is, to test for the equivalence of some but not all factor loadings and thresholds (Byrne, Shavelson, & Muthén, 1989). Previous MI tests of political trust have commonly opted for this solution (see Table 2). Especially in large-N studies, however, discerning patterns in modification indices to determine which parameters should be estimated freely becomes increasingly unwieldy (Byrne & van de Vijver, 2010).

Another, hitherto unexplored alternative to this data-driven solution in MI tests of political trust is a theory-driven strategy. Byrne and van de Vijver (2010) suggest testing the MI of subsamples of countries clustered according to a theoreti-

9 Chile, Colombia, Lebanon, Malaysia, Mexico, Peru, South Africa, and Taiwan.

10 This cross-loading was also non-significant in Lebanon.

11 In addition, in Model C2 the residual covariance matrix was not positive definite in Japan. The residual variance for trust in government was negative, indicating that the estimated factor loading did not fit the data well.

cally meaningful criterion. With regard to political trust, the post-communist countries are a case in point. Shortly after the end of the Cold War, Mishler and Rose (1994) argued that citizens in these countries cannot clearly distinguish between political trustees because they lack experience with them. From the perspective of political socialization theory, one could argue that almost three decades of democratic socialization have refined, and possibly diversified, people's construct of political trust in former communist countries in Europe more (Klingemann, Fuchs, & Zielonka, 2006). Inspired by these arguments, the MI of political trust was tested for the subsample of six post-communist European democracies in this study (Model C3). Full invariance of the model was supported by the data from Poland, Romania, and Slovenia. These results indicate that Mishler and Rose's (1994) general verdict no longer holds.[12] What is more, this brief demonstration of a theory-driven strategy to establish MI shows that similar tests for other subsets of countries could add to our insights on existing theoretical assumptions about the reasons for MI of political trust or lack thereof.

Insights and Recommendations for Future Political Trust Research

This article set out to answer to what extent the MI of political trust can be established in 32 democracies across the globe by means of MGCFA and if so, based on which measurement model. The single-country analyses showed that the data supported the two-dimensional model of trust in implementing and representative political institutions best. In the MGCFA, this model was not equivalent across all 32 democracies, however, because of three sources of bias (van de Vijver & Tanzer, 2004). First, item bias of 'trust in civil service' affected the model fit. Second, construct bias was apparent: The latent factor of trust in representative institutions did not sufficiently account for the shared variance between trust in parliament and trust in political parties in all countries. 'Trust in civil service' was therefore dropped and an error covariance was added to the measurement model in order to measure the construct of political trust in a more valid manner. Configural invariance of this revised two-dimensional model was established in 19 democracies. Additional revisions may be required in order to successfully remedy construct bias in the remaining 13 countries. Third, while the revised measurement model was fully invariant in three post-communist countries in eastern and southeastern Europe, the results suggest that method bias prevented full invariance in the other countries. Non-invariance of factor loadings and the thresholds indicate that the respondents did not use the response scale in the same manner.

12 See Schaap and Scheepers (2014) for a similar finding.

These results support previous studies and contradict others. They are in line with authors who distinguish between political trust in implementing and representative institutions conceptually (see for example Gabriel, 1999). Likewise, the analysis corroborates those empirical studies that found political trust to be two-dimensional (see Tables 1 and 2). Like previous analyses (see for example Braun, 2013 in Table 1), it also empirically reflects the ambiguity of the position of trust in civil service in the two dimensions of political trust described at the beginning of the article. The study does not, however, correspond to MGCFA that established MI of a single-dimensional model of political trust in Europe. This may be because the items used were not identical.

The results of this study underline that measurement invariance of political trust must not be assumed when testing theories about the changes, sources or consequences of political trust. Comparative political trust researchers can enhance the validity of their research findings on the generalizability of political trust theories by specifying the measurement model appropriately and carefully selecting the political trust items and countries. The findings therefore remind comparative researchers to use the ample cross-national survey data available methodically.

The findings are also informative for the future conceptualization of political trust. They provide reason to infer that, by and large, people in democracies across the globe have a two-dimensional construct of political trust. More conceptual work is needed, however, to identify the pertinent political trustees within these dimensions across countries.

In addition, the study contributes to insights regarding the valid measurement of political trust. Because the item 'trust in civil service' is apparently not as meaningfully related to the construct of political trust as the other items, future studies should carefully consider whether to include it. On a more general note, the study criticized the fact that there is no common set of comparable items to measure political trust. Such a set is crucial, however, because the content of the measured construct may be altered depending on the chosen items (Byrne & van de Vijver, 2010). Lack thereof impedes the cumulation of research on political trust.

A number of questions follow from this study. Future comparative research on political trust could study the reasons for the apparent bias. Do country-specific response tendencies affect MI and if so, why do they occur with items of political trust? Why is it so difficult to measure civil service in a comparable manner across countries? Last but not least, the study raises questions about the sources of political trust. The error covariance between trust in parliament and political parties indicates that they are not exclusively determined by people's overall level of trust. This could imply that their sources are more trustee-specific than those of the overall construct of political trust. Overall, the results of the study suggest that, in democracies, political trust is neither a single-dimensional construct nor a blanket judgment.

References

André, S. (2014). Does trust mean the same for migrants and natives? Testing measurement models of political trust with multi-group confirmatory factor analysis. *Social Indicators Research*, *115*, 963–982. doi.org/10.1007/s11205-013-0246-6

Ariely, G. (2015). Trusting the press and political trust: A conditional relationship. *Journal of Elections, Public Opinion and Parties*, *25*(3), 351–367. doi.org/10.1080/17457289.2014.997739

Ariely, G., & Davidov, E. (2012). Assessment of measurement equivalence with cross-national and longitudinal surveys in political science. *European Political Science*, *11*, 363–377. doi.org/10.1057/eps.2011.11

Asparouhov, T., & Muthén, B. O. (2010). *Weighted least squares estimation with missing data (*Mplus Technical Appendix). Retrieved from Mplus website: https://www.statmodel.com/download/GstrucMissingRevision.pdf

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186–203. dx.doi.org/10.1207/s15328007sem1302_2

Bowen, N. K., & Masa, R. D. (2015). Conducting measurement invariance tests with ordinal data: A guide for social work researchers. *Journal of the Society for Social Work and Research*, *6*(2), 229–249. doi.org/10.1086/681607

Braun, D. (2013). *Politisches Vertrauen in neuen Demokratien*. Wiesbaden, Germany: Springer VS.

Breustedt, W., & Stark, T. (2015). Thinking outside the democratic box. Political values, performance and political support in authoritarian regimes: A comparative analysis. In C. M. Eder, I. Mochmann, & M. Quandt (Eds.), *Political trust and disenchantment with politics: International perspectives* (pp. 184–222). Leiden, Netherlands: Brill.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. doi.org/10.1037/0033-2909.105.3.456

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*(2), 107–132. doi.org/10.1080/15305051003637306

Catterberg, G., & Moreno, A. (2006). The individual bases of political trust: Trends in new and established democracies. *International Journal of Public Opinion Research*, *18*(1), 31–48. doi.org/10.1093/ijpor/edh081

Center for Systemic Peace. (2017). *Polity IV Annual Time-Series 1800-2016 (p4v2016)* [Data file]. Retrieved from http://www.systemicpeace.org/inscrdata.html

Chang, E. C. C., & Chu, Y. (2006). Corruption and trust: Exceptionalism in Asian democracies? *The Journal of Politics*, *68*(2), 259–271. doi.org/10.1111/j.1468-2508.2006.00404.x

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. doi.org/10.1037/a0013193

Coromina, L., & Davidov, E. (2013). Evaluating measurement invariance for social and political trust in western Europe over four measurement time points (2002-2008). *Ask. Research and Methods*, *22*(1), 37–54.

Dalton, R. J. (2004). *Democratic challenges, democratic choices: The erosion of political support in advanced industrial democracies*. Oxford, England: Oxford University Press.

Davidov, E., Datler, G., Schmidt, P., & Schwartz, S. H. (2011). Testing the invariance of values in the benelux countries with the European Social Survey: Accounting for ordinality. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 149–171). New York, NY: Routledge.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*(4), 558–575. doi.org/10.1177/0022022112438397

Davidov, E., Meulemann, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75. doi.org/10.1146/annurev-soc-071913-043137

Denters, B., Gabriel, O. W., & Torcal, M. (2007). Political confidence in representative democracies: Socio-cultural vs. political explanations. In J. W. van Deth, J. R. Montero, & A. Westholm (Eds.), *Citizenship and involvement in European democracies: A comparative analysis* (pp. 66–87). London, England: Routledge.

Easton, D. (1975). A re-assessment of the concept of political support. *British Journal of Political Science*, *5*(4), 435–457. doi.org/10.1017/S0007123400008309

Fisher, J., van Heerde, J., & Tucker, A. (2010). Does one trust judgement fit all? Linking theory and empirics. *British Journal of Politics & International Relations*, *12*(2), 161–188. doi.org/10.1111/j.1467-856X.2009.00401.x

Fuchs, D., Gabriel, O. W., & Völkl, K. (2002). Vertrauen in politische Institutionen und politische Unterstützung. Österreichische Zeitschrift für Politikwissenschaft, *31*(4), 427–450.

Gabriel, O. W. (1999). Integration durch Institutionenvertrauen? Struktur und Entwicklung des Verhältnisses der Bevölkerung zum Parteienstaat und zum Rechtsstaat im vereinigten Deutschland. In J. Friedrichs & W. Jagodzinski (Eds.), *Soziale Integration* (pp. 199–235). Wiesbaden, Germany: Westdeutscher Verlag.

Hardin, R. (2000). The public trust. In S. J. Pharr & R. D. Putnam (Eds.), *Disaffected democracies: What's troubling the trilateral countries?* (pp. 31–51). Princeton, NJ: Princeton University Press.

Hooghe, M. (2011). Why there is basically only one form of political trust. *British Journal of Politics and International Relations*, *13(2)*, 269–275. doi.org/10.1111/j.1467-856X.2010.00447.x

Hooghe, M., & Kern, A. (2015). Party membership and closeness and the development of trust in political institutions: An analysis of the European Social Survey, 2002-2010. *Party Politics*, *21*(6), 944–956. doi.org/10.1177/1354068813509519

Horn, J. L., & Mcardle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. doi.org/10.1080/03610739208253916

Jackman, S. (2008). Measurement. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 119–151). Oxford, England: Oxford University Press.

Kankaraš, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research*, *40*(2), 279–310. doi.org/10.1177/0049124111405301

Kline, R. B. (2016). *Principles and practices of structural equation modeling* (4th ed.). New York, NY: The Guilford Press.

Klingemann, H.-D., Fuchs, D., & Zielonka, J. (Eds.). (2006). *Democracy and political culture in eastern Europe*. New York, NY: Routledge.

Levi, M., & Stoker, L. (2000). Political trust and trustworthiness. *Annual Review of Political Science*, *3*, 475–507. doi.org/10.1146/annurev.polisci.3.1.475

Listhaug, O., & Ringdal, K. (2008). Trust in political institutions. In H. Ervasti, T. Fridberg, M. Hjerm, & K. Ringdal (Eds.), *Nordic social attitudes in a European perspective* (pp. 131–151). Cheltenham, England: Edward Elgar.

Lu, P. (2014). A comparative analysis of political confidence in the BRICS countries. *Japanese Journal of Political Science*, *15*(3), 417–441. doi.org/10.1017/S1468109914000176

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*(4), 514–534. doi.org/10.1207/s15328007sem1104_2

Lühiste, K. (2006). Explaining trust in political institutions: Some illustrations from the Baltic states. *Communist and Post-Communist Studies*, *39*(4), 475–496. doi.org/10.1016/j.postcomstud.2006.09.001

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. doi.org/10.1037//1082-989X.1.2.130

March, J. G., & Olsen, J. P. (1989). *Rediscovering institutions: The organizational basis of politics*. New York, NY: The Free Press.

Marien, S. (2011a). The effect of electoral outcomes on political trust: A multi-level analysis of 23 countries. *Electoral Studies*, *30*, 712–726. doi.org/10.1016/j.electstud.2011.06.015

Marien, S. (2011b). Measuring political trust across time and space. In S. Zmerli & M. Hooghe (Eds.), *Political trust: Why context matters* (pp. 13–46). Colchester, England: ECPR Press.

Marien, S. (2017). The measurement equivalence of political trust. In S. Zmerli & T. W. G. van der Meer (Eds.), *Handbook on political trust* (pp. 89–103). Cheltenham, England: Edward Elgar.

Marshall, M. G., & Gurr, T. R. (2014). *Polity IV Project*. Retrieved from http://www.systemicpeace.org/polity/keynew.htm

Marshall, M. G., Gurr, T. R., & Jaggers, K. (2015). *Polity IV Project: Regime characteristics and transitions, 1800-2015. Dataset users' manual*. Vienna, Austria: Center for Systemic Peace. Retrieved from http://systemicpeace.org/inscr/p4manualv2015.pdf

Meade, A., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. doi.org/10.1037/0021-9010.93.3.568

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, *3*(1), 111–121. dx.doi.org/10.21500/20112084.857

Miller, A. H., & Listhaug, O. (1990). Political parties and confidence in government: A comparison of Norway, Sweden and the United States. *British Journal of Political Science*, *20*(3), 357–386. doi.org/10.1017/S0007123400005883

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515. dx.doi.org/10.1207/S15327906MBR3903_4

Mishler, W., & Rose, R. (1994). Support for parliaments and regimes in the transition toward democracy in eastern Europe. *Legislative Studies Quarterly*, *19*(1), 5–32. Retrieved from http://www.jstor.org/stable/439797

Mishler, W., & Rose, R. (1997). Trust, distrust and skepticism: Popular evaluations of civil and political institutions in post-communist societies. *The Journal of Politics*, *59*(2), 418–451. Retrieved from http://www.jstor.org/stable/2998171

Mishler, W., & Rose, R. (2001). What are the origins of political trust? Testing institutional and cultural theories in post-Communist societies. *Comparative Political Studies*, *34*(1), 30–62. doi.org/10.1177/0010414001034001002

Möllering, G. (2006). Trust, institutions, agency: Towards a neoinstitutional theory of trust. In R. Bachmann & A. Zaheer (Eds.), *Handbook of trust research* (pp. 355–376). Cheltenham, England: Edward Elgar.

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (Mplus Web Notes 4). Retrieved from Mplus website: https://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus: Statistical analysis with latent variables. User's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Newton, K., & Zmerli, S. (2011). Three forms of trust and their association. *European Political Science Review, 3*(2), 169-200. doi.org/10.1017/S1755773910000330

Norris, P. (2011). *Democratic deficit: Critical citizens revisited*. Cambridge, England: Cambridge University Press.

Oskarsson, S. (2010). Generalized trust and political support: A cross-national investigation. *Acta Politica*, *45*(4), 423–443. doi.org/10.1057/ap.2010.3

Pendergast, L. L., von der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology*, *60*, 65–82. doi.org/10.1016/j.jsp.2016.11.002

Pickel, G., & Walz, D. (1995). Politisches Institutionenvertrauen in der Bundesrepublik Deutschland in zeitlicher Perspektive. *Journal für Sozialforschung*, *35*(2), 145–155.

Poznyak, D., Meulemann, B., Abts, K., & Bishop, G. F. (2014). Trust in American government. Longitudinal measurement equivalence in the ANES, 1964-2008. *Social Indicators Research*, *118*, 741–758. doi.org/10.1007/s11205-013-0441-5

Rivetti, P., & Cavatorta, F. (2017). Functions of political trust in authoritarian settings. In S. Zmerli & T. van der Meer (Eds.), *Handbook on political trust* (pp. 53–68). Cheltenham, England: Edward Elgar.

Rogge, J., & Kittel, B. (2014). Politisches Vertrauen in Europa: Das Zusammenwirken von Demokratiequalität und Korruption. *Zeitschrift für Vergleichende Politikwissenschaft*, *8*(2), 155–178. doi.org/10.1007/s12286-014-0202-0

Rose, R., & Mishler, W. (2010). *Political trust and distrust in post-authoritarian contexts*. Aberdeen, Scotland: Center for the Study of Public Policy, University of Aberdeen.

Rothstein, B., & Stolle, D. (2003). Social capital, impartiality and the welfare state: An institutional approach. In M. Hooghe & D. Stolle (Eds.), *Generating social capital: Civil society and institutions in comparative perspective* (pp. 191–209). Basingstoke, England: Palgrave Macmillan.

Rothstein, B., & Stolle, D. (2008). Political institutions and generalized trust. In D. Castiglione, J. W. van Deth, & G. Wolleb (Eds.), *The handbook of social capital* (pp. 273–302). New York, NY: Oxford University Press.

Schaap, D., & Scheepers, P. (2014). Comparing citizens' trust in the police across European countries: An assessment of cross-country measurement equivalence. *International Criminal Justice Review*, *24*(1), 82–98. doi.org/10.1177/1057567714524055

Schneider, I. (2017). Can we trust measures of political trust? Assessing measurement equivalence in diverse regime types. *Social Indicators Research, 133*(3), 963-984. doi.org/10.1007/s11205-016-1400-8

Slomczynski, K. M., & Janicka, K. (2009). Structural determinants of trust in public institutions: Cross-national differentiation. *International Journal of Sociology*, *39*(1), 8–29. doi.org/10.2753/IJS0020-7659390101

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, England: Cambridge University Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. doi.org/10.1177/109442810031002

van Deth, J. W. (2013). Equivalence in comparative research: Staying in the middle of the road. In J. W. van Deth (Ed.), *Comparative politics: The problem of equivalence* (pp. xiii–xxvii). Colchester, England: ECPR Press.

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119–135. doi.org/10.1016/j.erap.2003.12.004

World Values Survey. (2017). *Who we are*. Retrieved from http://www.worldvaluessurvey.org/WVSContents.jsp

World Values Survey Association. (2016). *World Values Survey Wave 6 2010-2014 (v.20150418)* [Data file]. Retrieved from http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Doctoral dissertation, University of California, Los Angeles). Retrieved from https://www.statmodel.com/download/Yudissertation.pdf

Zmerli, S. (2004). Politisches Vertrauen und Unterstützung. In J. W. van Deth (Ed.), *Deutschland in Europa: Ergebnisse des European Social Survey 2002-2003* (pp. 229–255). Wiesbaden, Germany: VS Verlag.

Zmerli, S. (2013). Social structure and political trust in Europe: Mapping contextual preconditions of a relational concept. In S. I. Keil & O. W. Gabriel (Eds.), *Society and democracy in Europe* (pp. 111–138). London, England: Routledge.

Zmerli, S., & Newton, K. (2008). Social trust and attitudes toward democracy. *Public Opinion Quarterly*, *72*(4), 706–724. doi.org/10.1093/poq/nfn054

Zmerli, S., Newton, K., & Montero, J. R. (2007). Trust in people, confidence in political institutions, and satisfaction with democracy. In J. W. van Deth, J. R. Montero, & A. Westholm (Eds.), *Citizenship and involvement in European democracies: A comparative analysis* (pp. 35–65). London, England: Routledge.

Zmerli, S., & van der Meer, T. (Eds.). (2017). *Handbook on political trust*. Cheltenham, England: Edward Elgar.

Appendix

Table A1 Country-Specific Sample Sizes and Missings per Item

country	n	item (trust in)	missings	percent missing	country	n	item (trust in)	missings	percent missing	country	n	item (trust in)	missings	percent missing
Argentina	1030	police	13	1.3	Australia	1477	police	25	1.7	Brazil	1486	police	6	0.4
		court	17	1.7			court	37	2.5			court	5	0.3
		government	20	1.9			government	29	2.0			government	15	1.0
		pol. parties	36	3.5			pol. parties	31	2.1			pol. parties	16	1.1
		parliament	46	4.5			parliament	38	2.6			parliament	35	2.4
		civil service	40	3.9			civil service	36	2.4			civil service	12	0.8
Chile	1000	police	11	1.1	Colombia	1512	police	5	0.3	Cyprus	1000	police	4	0.4
		court	15	1.5			court	23	1.5			court	17	1.7
		government	15	1.5			government	12	0.8			government	18	1.8
		pol. parties	14	1.4			pol. parties	18	1.2			pol. parties	14	1.4
		parliament	22	2.2			parliament	26	1.7			parliament	13	1.3
		civil service	30	3.0			civil service	14	0.9			civil service	10	1.0
Estonia	1533	police	14	0.9	Georgia	1202	police	40	3.3	Germany	2046	police	21	1.0
		court	54	3.5			court	111	9.2			court	53	2.6
		government	21	1.4			government	53	4.4			government	45	2.2
		pol. parties	60	3.9			pol. parties	58	4.8			pol. parties	64	3.1
		parliament	41	2.7			parliament	58	4.8			parliament	68	3.3
		civil service	45	2.9			civil service	54	4.5			civil service	44	2.2

Table A1 continued

country	n	item (trust in)	missings	percent missing	country	n	item (trust in)	missings	percent missing	country	n	item (trust in)	missings	percent missing
Ghana	1552	police	0	0.0	India	1581	police	4	0.3	Japan	2443	police	144	5.9
		court	0	0.0			court	3	0.2			court	254	10.4
		government	0	0.0			government	4	0.3			government	277	11.3
		pol. parties	0	0.0			pol. parties	4	0.3			pol. parties	333	13.6
		parliament	0	0.0			parliament	4	0.3			parliament	322	13.2
		civil service	0	0.0			civil service	4	0.3			civil service	338	13.8
Lebanon	1200	police	43	3.6	Malaysia	1300	police	1	0.1	Mexico	2000	police	1	0.05
		court	57	4.7			court	1	0.1			court	20	1.0
		government	221	18.4			government	2	0.2			government	5	0.2
		pol. parties	80	6.7			pol. parties	2	0.2			pol. parties	3	0.1
		parliament	85	7.1			parliament	1	0.1			parliament	25	1.2
		civil service	53	4.4			civil service	1	0.1			civil service	29	1.4
Netherlands	1902	police	63	3.3	New Zealand	841	police	39	4.6	Peru	1210	police	7	0.6
		court	78	4.1			court	55	6.5			court	17	1.4
		government	89	4.7			government	81	9.6			government	22	1.8
		pol. parties	102	5.4			pol. parties	73	8.7			pol. parties	29	2.4
		parliament	133	7.0			parliament	76	9.0			parliament	14	1.2
		civil service	132	6.9			civil service	111	13.2			civil service	22	1.8
Philippines	1200	police	1	0.1	Poland	966	police	50	5.2	Romania	1503	police	27	1.8
		court	2	0.2			court	80	8.3			court	91	6.1
		government	2	0.2			government	38	3.9			government	50	3.3
		pol. parties	0	0.0			pol. parties	60	6.2			pol. parties	65	4.3
		parliament	1	0.1			parliament	56	5.8			parliament	62	4.1
		civil service	1	0.1			civil service	73	7.6			civil service	57	3.8

Table A1 continued

country	n	item (trust in)	missings	percent missing	country	n	item (trust in)	missings	percent missing	country	n	item (trust in)	missings	percent missing
Slovenia	1069	police	22	2.1	South Africa	3531	police	99	2.8	South Korea	1200	police	4	0.3
		court	50	4.7			court	129	3.7			court	4	0.3
		government	29	2.7			government	112	3.2			government	3	0.2
		pol. parties	31	2.9			pol. parties	128	3.6			pol. parties	6	0.5
		parliament	26	2.4			parliament	132	3.7			parliament	6	0.5
		civil service	32	3.0			civil service	193	5.5			civil service	4	0.3
Spain	1189	police	17	1.4	Sweden	1206	police	7	0.6	Taiwan	1238	police	50	4.0
		court	24	2.0			court	34	2.8			court	87	7.0
		government	18	1.5			government	21	1.7			government	68	5.5
		pol. parties	25	2.1			pol. parties	34	2.8			pol. parties	90	7.3
		parliament	55	4.6			parliament	32	2.7			parliament	96	7.8
		civil service	42	3.5			civil service	220	18.2			civil service	69	5.6
Trinidad and Tobago	999	police	24	2.4	Turkey	1605	police	21	1.3	Ukraine	1500	police	0	0.0
		court	75	7.5			court	47	2.9			court	0	0.0
		government	48	4.8			government	41	2.6			government	0	0.0
		pol. parties	56	5.6			pol. parties	52	3.2			pol. parties	0	0.0
		parliament	69	6.9			parliament	62	3.9			parliament	0	0.0
		civil service	106	10.6			civil service	64	4.0			civil service	0	0.0

Table A1 continued

country	n	item (trust in)	missings	percent missing	country	n	item (trust in)	missings	percent missing	country	n	item (trust in)	missings	percent missing
United States	2232	police	37	1.7	Uruguay	1000	police	18	1.8					
		court	44	2.0			court	57	5.7					
		government	45	2.0			government	33	3.3					
		pol. parties	44	2.0			pol. parties	53	5.3					
		parliament	62	2.8			parliament	64	6.4					
		civil service	50	2.2			civil service	96	9.6					

Note. Data are from the World Values Survey 2010-2012. Own compilation.

Table A2 Comparison of Configural Invariance Results with Different Reference Indicators for Model A

reference indicator	χ^2 (df)	p-value	CFI	TLI	RMSEA (90% CI)
trust in parliament and trust in police	6457.907 (256)	0.00	0.987	0.976	0.129 (0.127-0.132)
trust in parliament and trust in court	6481.266 (256)	0.00	0.987	0.976	0.130 (0.127-0.132)
trust in political parties and trust in police	6453.700 (256)	0.00	0.987	0.976	0.129 (0.127-0.132)
trust in political parties and trust in court	6471.272 (256)	0.00	0.987	0.976	0.130 (0.127-0.132)
trust in government and trust in police	6454.196 (256)	0.00	0.987	0.976	0.129 (0.127-0.132)
trust in government and trust in court	6485.580 (256)	0.00	0.987	0.976	0.130 (0.127-0.132)
trust in civil service and trust in police	6459.617 (256)	0.00	0.987	0.976	0.129 (0.127-0.132)
trust in civil service and trust in court	6490.506 (256)	0.00	0.987	0.976	0.130 (0.127-0.132)
factor variance=1/factor mean=0	6457.732 (256)	0.00	0.987	0.976	0.129 (0.127-0.132)

Note. WLSMV estimator (theta parameterization), pairwise present was used to handle missing data (Asparouhov & Muthén, 2010), df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis-Index, RMSEA = root mean square error of approximation, 90% CI = 90% confidence interval. Data are from the World Values Survey 2010-2012, 32 countries.

methods, data, analyses | Vol. 12(1), 2018, pp. 47-76 DOI: 10.12758/mda.2017.11

Testing Measurement Invariance for a Second-Order Factor. A Cross-National Test of the Alienation Scale

Maksim Rudnev[1], *Ekaterina Lytkina*[1], *Eldad Davidov*[2], *Peter Schmidt*[3] & *Andreas Zick*[4]

[1] *National Research University Higher School of Economics,*
[2] *University of Cologne and University of Zurich,*
[3] *University of Giessen,* [4] *University of Bielefeld*

Abstract

Multiple group confirmatory factor analysis has become the most common technique for assessing measurement invariance. However, higher-order factor modeling is less frequently discussed in this context. In particular, the literature provides only very general guidelines for testing measurement invariance of second-order factor models, which is a prerequisite for conducting meaningful comparative research using higher-order factors. The current paper attempts to fill this gap. First, we explicate the constraints required for identification of the invariance levels in a multiple group second-order factor model. Second, in addition to the conventional interpretation of the results of this assessment, we suggest an alternative view on the invariance properties of a second-order factor as evidence of structural rather than measurement invariance. Third, we present an empirical application of the test which builds on Seeman's alienation scale and utilizes data from eight countries collected in 2008-2009. We found empirical support for metric invariance of both the first- and second-order factors, but no support for scalar invariance of the first- and second-order factors. However, we find pairs of countries where scalar invariance for both the first- and second-order factors is supported by the data. We finalize with a discussion of the results and their interpretation.

Keywords: higher-order factors, measurement invariance, multiple group confirmatory factor analysis, anomie, Seeman's alienation scale

Introduction*

Measurement invariance is the degree to which the measurement model of a latent variable is the same across groups involved in the analysis. It is considered to be one important indicator of population homogeneity. In recent years, various studies have emphasized that the assessment of measurement invariance is necessary in studies involving latent variables and multiple samples, especially in cross-national survey research (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Davidov, Schmidt, & Billiet, 2011). There are several approaches to assess measurement invariance of latent variables; these include lenient ones such as multidimensional scaling and exploratory factor analysis, and stricter ones such as multiple group confirmatory factor analysis (MGCFA: Jöreskog, 1971) or multiple group latent class analysis (McCutcheon, 1987). Since its introduction for the assessment of measurement invariance (Meredith, 1993), MGCFA has become very popular (Davidov et al., 2014) as demonstrated by its inclusion in numerous textbooks and statistical guides, with hundreds of published papers demonstrating its applicability for invariance testing.

Different extensions of the basic MGCFA model have also been discussed in the literature. However, one variant of the MGCFA model, namely, its application to second-order and higher-order factor models, has received considerably less attention. A second-order factor model implies an ordinary factor model in which covariances of latent variables (i.e. first-order factors) are determined by one or more higher-order latent variables (i.e. second-order factors, see Figure 1). In cases of three or more second-order factors, third-order factor models are possible, although such models are rarely used (for an exception, see e.g., Cieciuch, Davidov, Vecchione, & Schwartz, 2014).

---

* This article is dedicated to Melvin Seeman of UCLA, the pioneer of theoretically driven empirical alienation research, in honor of his 100 birthday on February 5, 2018! He is still going strong in his work on the topic and is now focusing on alienation and health.

Acknowledgments

Work of the first author is the outcome of the Basic research program of National Research University Higher School of Economics. Work of the second author was funded by the Russian Academic Excellence Project '5-100'. The third author would like to thank the University of Zurich Research Priority Program "Social Networks" for their financial support. The fourth author's work was supported by a Humboldt fellowship of the Polish Foundation for basic research. All authors would like to thank Lisa Trierweiler for the English proof of the manuscript.

Direct correspondence to

Maksim Rudnev

E-mail: mrudnev@hse.ru

Measurement models with second-order factors are good representations of second-order concepts (Rindskopf & Rose, 1988). For example, the popular Big Five personality traits structure (Costa & McCrae, 1990) was modeled as a set of second-order factors of Cattell's 16 first-order factors (John & Srivastava, 1999). A general intelligence, or Spearman's g-factor, can similarly be seen as a second-order factor where verbal, mathematical, and other kinds of intellectual abilities act as first-order factors (Jensen, 1998). Basic human values are structured hierarchically as well: There are specific values and higher-order values (Schwartz et al., 2012). Finally, alienation can be expressed as a higher-order concept for powerlessness, meaninglessness, and isolation (Seeman, 1991). We will go into more detail about this concept below in the empirical part of the study.

A second-order factor model mimics the logic of the first-order factor models. First-order factor models represent the reflective relations between observed indicators and an underlying factor (latent variable) (Boorsbom, Mellenbergh, & van Heerden, 2003; Costner, 1969; Hempel, 1973). Similarly, second-order factor models represent the reflective relations between first-order factors and an underlying second-order factor (which is also a latent variable). However, when it comes to testing the measurement invariance of second-order factors in multiple groups, various complications occur. Despite the growing number of substantive papers (over 500)[1] addressing second-order factor measurement invariance, very few of these attempted to describe the strategies and complications of this method. Chen, Sousa, and West (2005) provided general guidelines for testing measurement invariance of second-order factor models. Dimitrov (2010) followed their approach and presented an empirical example using the software package Mplus (Muthén & Muthén, 1998-2016). Strasheim (2011) explicated this approach using matrix notation and supplemented it with a technical description of the possible levels of measurement invariance for second-order factors, including the ones that are rarely used (e.g., invariance of residuals).

The purpose of the current paper is twofold. First, we provide a simple, non-technical yet comprehensive description of procedures involved in the assessment of measurement invariance of second-order factor models, embedding these into the context of cross-country surveys. Second, we demonstrate the procedure on real data and test for measurement invariance of a second-order factor. This second-order factor represents alienation, an important concept in sociological literature (Seeman, 1983). We test its measurement invariance properties across eight countries. Thus, rather than presenting a novel procedure, the added value of the paper focuses on guiding the reader through the process of assessing measurement invariance of second-order factors, providing a step-by-step description of the procedure, implementing the method on data across a number of countries, and presenting the

1 This is the number of papers citing Chen et al. (2005) paper in Google Scholar as of February 25, 2017, most of which test second-order factor invariance in some form.

example codes. Furthermore, we suggest an alternative interpretation of second-order factors across groups as a manifestation of structural rather than measurement parameters.

In the next section, we first describe different hierarchical levels of measurement invariance tests and how they apply to second-order factor models. Next, we discuss identification issues and different possible interpretations of the hierarchical factor structure. Finally, we present a cross-national measurement invariance test of alienation in a second-order multigroup factor model.

Assessment of Measurement Invariance

Measurement Invariance of First-Order Factor Models

A common way to assess measurement invariance is to specify an MGCFA model across groups, such as countries, cultures, language groups, or any other nominal variable (Davidov et al., 2014). MGCFA models are fitted to the data using different sets of specific constraints that correspond to the specific level of measurement invariance. Researchers typically differentiate between three levels of measurement invariance that are sufficient for conducting most comparative survey data analyses: configural, metric, and scalar invariance (Vandenberg & Lance, 2000; but see, e.g., Meredith, 1993, for additional levels of invariance).

Configural invariance means that approximately the same concept is measured across groups. It does not guarantee that a construct is measured on the same scale with the same zero point, but it indicates whether higher factor values correspond to higher levels of a concept measured in several groups. Support for configural invariance allows meaningful between-group comparison of *signs* of correlations or regression coefficients, which describe association of the latent variable with exogenous (i.e., external to the measurement model) variables. Configural invariance is met when the general factor structure is the same across groups, including the number of factors and the general pattern of factor loadings. Testing for configural invariance does not involve any parameter constraints across groups except those required for model identification (discussed below). Therefore, configural invariance may also be assessed with "lenient" methods, including multidimensional scaling (e.g., Schwartz & Bilsky, 1990) or exploratory factor analysis (Horn & McArdle, 1992; Lorenzo-Seva & Ten Berge, 2006). These methods provide statistical criteria on the degree of similarity between factor loadings across groups but are not methods considered to be strict tests. Testing for higher levels of measurement invariance is strict albeit necessary when researchers are interested in comparisons of latent variables' degree of association or means across groups.

Metric invariance represents a second and higher level of measurement invariance. It means that the constructs are measured by the same measurement units across groups. Nevertheless, it does not guarantee that the zero point of the scales is the same across groups. Metric invariance implies that any difference in one unit of a latent variable results in the same differences of the observed indicator variables in all groups. It follows that when metric invariance is present, covariances and unstandardized regression coefficients involving latent variables can be meaningfully compared across groups. Metric invariance is met when the factor loadings are the same across groups. It is assessed by fixing factor loadings to be equal across groups and checking whether the model fit significantly decreases in comparison to the configural model.

Scalar invariance represents the third level of measurement invariance and means that the latent variables' scales are measured with the same units and have the same zero point for all the groups included in the analysis. It implies that the levels of the latent variables correspond to the same levels of the manifest variables across groups. Therefore, in addition to covariances and unstandardized regression coefficients, the means of the latent variables (the latent means) may be meaningfully compared across groups. Scalar invariance is met when intercepts of the observed indicator variables (in addition to the factor loadings) are the same across groups. Consequently, it is assessed by constraining the intercepts of the same items across different groups to equality.

One may rely on partial metric or partial scalar invariance in situations where not all the factor loadings and/or intercepts are the same across groups. Partial invariance would require at least two items with equal factor loadings (for partial metric invariance) and at least two items with equal factor loadings and intercepts per factor (for partial scalar invariance) to be invariant (Byrne, Shavelson, & Muthén, 1989). Partial metric or scalar invariance has the same implications as the corresponding full metric or full scalar invariance (but for criticisms on this approach, see, e.g., Steinmetz, 2011). Similarly, partial invariance may be applicable also for higher-order factors as discussed below.

Measurement Invariance of Second-Order Factor Models

Assessment of measurement invariance of second-order factor models follows basically the same logic as the assessment of measurement invariance of first-order models but with minor differences.

Before testing for measurement invariance of a second-order factor, it is necessary to establish invariance of the first-order factors. Metric invariance of the first-order factors is a prerequisite for the assessment of configural and metric invariance of the second-order factor. Scalar invariance of the first-order factors

is a prerequisite to assess scalar invariance of the second-order factor. This determines the sequence of the models when assessing measurement invariance.

The metric invariance model of the first-order factors serves as the model where configural invariance of the second-order factor is tested for the following reason: If metric invariance of the first-order factors is supported by the data, it implies that covariances between the first-order factors are comparable. Therefore, the loadings of the second-order factors can be meaningfully compared across groups. Researchers can then examine the second-order factor loadings to determine whether their structure is also similar across countries. This can be done by fixing the second-order loadings to equality across groups.

The model parameter constraints used to test for second-order scalar invariance are similar to those applied in testing for scalar invariance of first-order factors (see Table 1) with slight differences. To test for scalar invariance of the second-order factor, scalar invariance of the first-order factors is necessary. It will imply that the means of the first-order factors are comparable and one may meaningfully test if they can be constrained to equality across groups.

Partial invariance of a second-order factor model may also be tested if full metric or scalar invariance is not supported by the data for the second-order factor. Following Byrne et al.'s (1989) suggestions for assessing partial invariance of first-order factors, a similar logic may be applied to second-order factors. According to this logic, two invariant first-order factors (with equal loadings on the second-order factor and equal intercepts) may be sufficient for guaranteeing partial invariance of the second-order factor. As this suggestion of implementing the idea of partial invariance on second-order factors is rather new, it requires further exploration using simulation studies that do not only focus on first-order factors (e.g., de Beuckelaer & Swinnen, 2011) but also on partial invariance of second-order factors.

A point worth mentioning is that measurement invariance of higher-order (e.g., of third- or fourth-order) factors follows a similar logic as the one for testing measurement invariance of second-order factors, because factors are continuous on all levels. While metric invariance is a prerequisite for configural and metric invariance on the higher factor level, and scalar invariance is a prerequisite for scalar invariance on the next higher factor level, it may make sense to consider testing first for metric invariance on all factor levels before assessing scalar invariance. By doing so, a differentiation between covariance and mean structures can be achieved.[2]

2 We would also like to indicate that this paper does not consider invariance of errors (the so-called strict invariance) for two reasons. First, this test is rarely conducted in cross-national applied research (see, e.g., Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009). Second, equal errors across groups imply equal variances of their corresponding indicators or factors, a situation which is highly unlikely to occur.

Model Identification

Identification of a variance-covariance structure of the first-order factors may be achieved in three interchangeable ways (Little, Slegers, & Card, 2006): fixing the factor variances to 1, fixing the sum of the factor loadings to 1 ("effect coding"), or fixing one factor loading per factor to 1 ("marker indicator"). Likewise, models with a mean structure can be identified either by fixing one intercept per factor to 0, the latent mean in one group to 0, or the sum of the intercepts to 0.

These identification methods differ in their suitability for measurement invariance testing. There is no reason to assume that variances of latent variables should be equal across groups when testing for configural, metric, or scalar invariance. Therefore, it may be problematic to fix factor variances to 1 in all groups. Constraining the sum of factor loadings to be equal across groups makes it difficult to detect model misspecifications, especially when some factor loadings differ across groups. Therefore, constraining one factor loading per factor to 1 is a preferred way of identification of the covariance part of first-order factors. A disadvantage of this approach is that, in the context of modeling multiple groups, this constraint implies equality of the corresponding parameter across groups; thus, a factor loading fixed to 1 is assumed to be invariant across groups a priori, even in the unconstrained configural model. If the fixed loading is in fact not invariant, other truly invariant loadings might be represented by noninvariant factor loading estimates to compensate for the misspecified model. Therefore, special attention should be paid to the selection of the indicator whose loading is fixed to 1. For example, one should try different marker indicators for identifying the model and examine the patterns of loading differences across groups. Researchers are recommended to choose the most reliable and invariant item to serve as a marker. Ideally, this item would also be conceptually closest to the latent variable underlying the different items. An improper selection of a marker variable may lead to incorrect detection of the invariance level when only partial invariance is given in the data (Johnson, Meade, & DuVernet, 2009; Jung & Yoon, 2017). A proper selection of the marker indicator would enable researchers to meaningfully interpret both factor loadings and latent means.

When testing for scalar invariance, the mean structure is easy to identify by constraining the first-order factor means in one reference group to 0. Another technique requires constraining the intercept of a reference indicator to 0 (the "marker indicator method"). We do not apply the former method, because the first-order factor means serve as (latent) intercepts for the second-order factors. When testing for scalar invariance of second-order factors, latent intercepts are constrained to be equal across groups, and being constrained to 0 in one group implies constraining them to zero in all the groups. It leads to the test of latent intercepts' being zero instead of desired test of their equality across groups. Therefore, when testing for

scalar invariance of second-order factors, we find it more appropriate to use the "marker method" by fixing one indicator intercept per first-order factor to 0. This allows first-order factor means to be freely estimated in all groups, and it is necessary for testing them for equality when assessing second- order factor scalar invariance.

When the latter method (i.e., the marker method) is used, an intercept fixed to 0 is assumed to be invariant across groups a priori, even in the unconstrained configural model, without empirically testing it. Just like with factor loadings, special attention should be paid to the selection of the indicator whose intercept is fixed to 0. For example, one may examine the modification indices to find out how the fit of the model would change if the marker indicator's intercept was not assumed to be invariant.

The identification of the second-order part of the model follows a similar rationale. For the variance-covariance structure one could either fix the second-order factor variance(s) or one of the second-order factor loading(s) to 1. Alternatively, one may fix the sum of the second-order factor loadings to 1 ("effect coding"). Also, in the context of group comparisons of second-order factors, it is not plausible to assume a cross-group equality of second-order factor variances. Therefore, a common way to identify the second-order part of the model is to choose one first-order factor to serve as an anchor and provide the metric for the second-order factor. Its loading to the second-order factor is constrained to 1. Again, attention should be paid to the selection of the metric, that is, the first-order factor, whose loading is fixed.

The means structure of the second-order factor may be identified by constraining the second-order factors' means in one group to 0. Alternatively, one may constrain the intercept of one reference ("marker") first-order factor to 0. We believe that identifying the second-order factor's mean by constraining it in one group to 0 is preferable and more convenient to implement, because its "indicators" (i.e., the first-order factors) are latent variables themselves whose means may be of interest for researchers. Consequently, it is reasonable to try to avoid constraining the intercept of one of them to 0 across groups.[3]

Testing Procedure

There are two strategies for testing these sequences of constraints. The top-down strategy requires first testing the most restrictive model, and then constraints are

3 When items are considered ordinal rather than continuous, in addition to factor loadings and intercepts one has to consider also a new type of parameters – thresholds (see, e.g., Davidov, Datler, Schmidt, & Schwartz, 2011). The issue of measurement invariance in the case of ordinal responses has not been fully clarified yet (see, e.g. Millsap, 2011, p. 129; Wu & Estabrook, 2016) and is beyond the scope of the current paper.

relaxed until an appropriate fit is achieved (Horn & McArdle, 1992). The bottom-up approach first tests the least restrictive models (i.e., configural invariance), and then factor loadings and intercepts are constrained in a stepwise manner. When working with second-order factor models, it is easier (and therefore preferable) to use the bottom-up strategy, because second-order factor models are complex and, in this way, it becomes easier to detect misspecifications (Brown, 2015, p. 290).

The sequence and specific sets of the constraints tested during the test for measurement invariance of the second-order factor models are listed in Table 1 and summarized below. First, configural invariance of the first-order factors is tested, followed by tests of first-order and second-order metric invariance. These are necessary preconditions to finally test for first- and second-order scalar invariance. Whereas tests of metric invariance on both levels require only information about the variance and covariance structure of the data, tests of scalar invariance on both levels require additional information on the mean structure of the data. Thus, one begins by testing for both first- and second-order metric invariance, and afterwards proceeds with testing for both types of scalar invariance. Such a sequence is reasonable because it allows differentiating in the invariance test between the covariance and the mean structures. Metric invariance on the second level is not a necessary requirement for scalar invariance on the first level. However, logically it makes sense to first examine whether metric invariance holds on both levels, and then expand the test using also information on the means and test for scalar invariance on both levels. As a general guideline, the logic of comparisons is not necessarily to choose the best-fitting model, but to select the most parsimonious one (i.e., the most constrained, with a highest possible level of invariance) which is still well-fitting (Brown, 2015). Such a model will allow more types of cross-group comparisons (as discussed previously). To achieve this, one can begin by comparing the fit of more constrained models with the less constrained ones. If the fit decreases considerably, we have to reject the model with a higher level of invariance, and if there is no considerable decrease in model fit, we can accept the model with a higher level of invariance.

What is a considerable decrease in model fit? The chi-square (χ^2) difference test (also known as the likelihood ratio test) is often applied to compare adjacent pairs of nested models, but it is known to reject models even when violations are minor, particularly when the sample size is large (Chen, 2007). Therefore, Chen (2007) and Cheung and Rensvold (2002) proposed to complement it with alternative criteria. They suggest that if the sample size is large, (>300), a comparative fit index (CFI) difference not larger than 0.01 across models implies that the model fit does not deteriorate considerably. In addition, one could use the sample-adjusted Bayesian information criterion (SABIC), whose values do not supply a significance level but are sensitive to measurement noninvariance; usually the most parsimonious yet well-fitting model has a lower SABIC (Van de Schoot, Lugtig, & Hox,

Table 1 Testing for Measurement Invariance and Possible Parameter Constraints in Multiple Group Confirmatory Factor Analysis with a Second-Order Factor

	First-order factors			Second-order factor	
	Factor loadings	Item intercepts	Latent means/ intercepts	Factor loadings	Latent means
1. Configural	Free, but one per factor is fixed to 1	Free, but one per factor is fixed to 0	Free	Free, but one per factor is fixed to 1	Fixed to 0
2. First-order metric	Set equal across groups and one per factor is fixed to 1	Free, but one per factor is fixed to 0	Free	Free, but one per factor is fixed to 1	Fixed to 0
3. First- and second-order metric	Set equal across groups and one per factor is fixed to 1	Free, but one per factor is fixed to 0	Free	Set equal across groups and one per factor is fixed to 1	Fixed to 0
4. First-order scalar	Set equal across groups and one per factor is fixed to 1	Set equal across groups and one per factor is fixed to 0	Free	Set equal across groups and one per factor is fixed to 1	Fixed to 0
5. First- and second-order scalar	Set equal across groups and one per factor is fixed to 1	Set equal across groups and one per factor is fixed to 0	Set equal across groups	Set equal across groups and one per factor is fixed to 1	Free, but fixed to 0 in one group

Note. The variances of all factors and residuals are freely estimated in all models. The models are based on the marker indicator approach (Little et al., 2006).

2012). Note that beside these criteria, the fit of each model should be acceptable on its own, that is, every model should fit the data well (but to a different degree). We consider a model fit as acceptable when the CFI value is at least as high as 0.90 (soft criterion) or 0.95 (very good fit), and the root mean square error of approximation (RMSEA) is not larger than 0.08 with the upper bound of its confidence interval not higher than 0.10 (but see, e.g., Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004, or West, Taylor, & Wu, 2012, for a vivid discussion on this topic). Thus, and given

that χ^2 testing leads too often to significant falsification, one may accept a model with a higher level of invariance if the model deterioration (e.g., in terms of CFI and RMSEA) is not too large and within the recommended criteria.

Interpreting Second-Order Factor Models in a Multiple Group Comparison

We suggest viewing measurement invariance of second-order factors using different perspectives. These perspectives rely on the two differing approaches on how to view second-order factors in the context of multiple group comparisons. The deductive and most popular approach assumes that the logic applied to first-order factor models (Costner, 1969; Hempel, 1973) should be transferred also to second-order factors (Chen et al., 2005; Dimitrov, 2010; Strasheim, 2011). From this point of view, scalar invariance for the second-order factor is necessary to compare its means across groups meaningfully.

The second interpretation originates from the realization of the fact that first-order factors are not observed variables; hence, they should not be treated in the same manner as indicators. Second-order factors might be treated as compensatory, that is, any combination of the invariant first-order factors is indicative of the general higher-order latent variable. The logic behind this view suggests that second-order factors based on invariant first-order factors reflect *structural* relations between the second- and the first-order factors rather than *measurement* relations. In other words, the relative importance of first-order factors may vary across societies or over time without changing the nature of the second-order factor. Thus, even if the structure (the relations between the first- and the second-order factors) slightly varies across groups, second-order factors may still be functionally equivalent across groups and could be compared (Hui & Triandis, 1985; Van de Vijver & Leung, 1997). Indeed, this view may be regarded as problematic, because strictly speaking, if measurement invariance of a second-order factor is not given, its means may be noncomparable. However, we believe it is worthwhile to consider the fact that the measurement structure of second-order factors may vary slightly across societies and over time even when they in fact tap into the very same general concept. One could take this into account by examining approximate (rather than exact) measurement invariance (Van de Schoot et al., 2013).[4]

4 An interesting alternative to the model with the single second-order factor is a bifactor model, which has a single factor loading on all of the items and has zero correlations with the other factors (Chen, West, & Sousa, 2006). Such a general factor might represent a method effect (e.g., response style) and can be easily confused with the second-order factor structure, especially in cross-national surveys. One can test the difference in fit of the second-order factor model and bifactor model, as they are nested, to determine which one represents the data better (Yung, Thissen, & McLeod, 1999).

This distinction corresponds to the difference between the etic and emic approaches in cross-cultural studies (Van de Vijver & Leung, 1997). Etic means that one postulates general statements which should hold in any culture, whereas the emic position assumes that relationships always vary depending on culture. Thus, etic corresponds to our first interpretation and emic to our second one. One should note, however, that this argument may also be used for interpreting the relation between items and first-order factors.

It may be of great interest to determine whether a higher-order construct has similar subdimensions with equal loadings across cultures or over time. This may be considered a major issue of investigation in comparative sociology for different types of concepts. Thus, when first-order factors display measurement invariance but second-order factors do not, it may not necessarily imply that the measurement of the items and their operationalization are problematic or that they are inadequate for comparative research. Instead, noninvariance of a second-order factor may imply that it has a different content across groups. Such an implication can be of great interest for theoreticians. In the following empirical example, we demonstrate how invariance of the second-order factor model is tested and interpreted.

Empirical Illustration

For the empirical illustration we use data measuring the concept of alienation, which is a concept of major importance in sociology. Initially defined by Karl Marx as "the surrender of control over work and its products, and the worker's disengagement from both work and fellow workers" (Seeman, 1991, p. 291), it denotes an individual's isolation, estrangement, and sense of being lost within the society (e.g., Seeman, 1959, 1991; see also Dean, 1961). The most stringent and also popular theoretical models of alienation were developed by Seeman (1959) who considered alienation as a combination of five subdimensions: feeling of powerlessness, meaninglessness, normlessness, isolation, and self-estrangement. A series of scales were developed based upon his model. Studies applying these scales connected the five subdimensions with the value-expectancy theory (see Robinson, 1973; Schmidt, 1990; Seeman, 1991) and applied them in several contexts (e.g., Dean, 1961; Huschka & Mau, 2006; McClosky & Schaar, 1965; Middleton, 1963). However, the validity and cross-national reliability of these scales have not been assessed yet (for an exception, see a German-American comparison of some of the items of the scales by Krebs & Schuessler, 1989). In addition, alienation was never specified and tested as a second-order factor model, although the underlying theoretical conceptualization would require this (Schmidt, 1990). Due to data constraints (see the next section), we employ and test the measurement of only three of the five subdimensions of alienation in the analysis. The definitions of the three subdimensions

are presented in Table 2. In the following section, we will test for measurement invariance of alienation using a shortened version of McClosky and Schaar's (1965) alienation scale across several European countries.

Data and Measures

We employ data from the project "Group-Focused Enmity" carried out in 2008/2009 by the Institute for Interdisciplinary Research on Conflict and Violence (Bielefeld University, Germany) with its European partners[5] in eight countries: France, Germany, Great Britain (England, Scotland, Wales, but not Northern Ireland), Hungary, Italy, the Netherlands, Poland, and Portugal. These countries were chosen because they represent old and new EU member states and different geographical regions in Europe (Küpper et al., 2010; Zick et al., 2011). The countries differ in various characteristics such as the level of economic prosperity, level of inequality, history of democracy, or their citizens' well-being. These features may contribute not only to different levels of alienation, but also to a different measurement structure of alienation. We expect countries with a longer history of democracy, longer EU membership, a stronger economy, and a higher level of democratic participation of citizens to have lower levels of alienation.

Data were collected via computer-assisted telephone interviews with a representative sample of about 1,000 respondents aged 16 years and above in each country. A representative random sample was drawn from the national telephone master samples (stratified according to a regional allocation of the population). After choosing a household, the target person was selected by either picking the household member whose birthday was next or last, or by the Kish grid method where a table of preassigned random numbers is used to choose a respondent (Kish, 1949). Response rates were rather low and varied across countries, ranging between 4.5% in Italy to 33% in Germany. In the final sample, 48% of respondents were male and 52% were female, and the mean age was 47 years. In each country sample, about 1,000 respondents were interviewed, but only about half of them were asked all the questions included in the scale. Thus, the actual sample size in each country used in our study was approximately 500 (see Appendix A). These samples do not differ systematically from the full samples in their sociodemographic characteristics such as age and gender. Missing values were handled with the full information maximum likelihood algorithm during model estimation (Arbuckle, 1996).

The alienation scale in the data included six indicators which measured three first-order concepts: powerlessness, meaninglessness, and social isolation. The

5 The project was financially supported by the Compagnia di San Paolo, the Freudenberg Stiftung, the Groeben Stiftung, the Volkswagen Stiftung, and two other private foundations. For further details on data collection and documentation, see Zick, Küpper, and Hövermann (2011) and Küpper, Wolf, and Zick (2010).

items represented a short version of the McClosky and Schaar (1965) scales, and their question wording is presented in Table 2. No measures for normlessness and self-estrangement were included in the data. However, we consider these three subdimensions of alienation, that is, powerlessness, meaninglessness, and social isolation, to be the very core of alienation (Dean, 1961, p. 754; Seeman, 1959, p. 787; Seeman, 1991, p. 339). All items were measured on an agree-disagree scale ranging from 1 to 4 and then recoded so that 1 indicated "strongly disagree" and 4 indicated "strongly agree." Alienation was modeled in each country sample as a second-order factor reflecting the three subdimensions, which were in turn measured by two items each (see Figure 1). The replication data are listed in Appendix E.

In the following section, we will explore whether scalar invariance of the alienation measurement model is given in the data. However, there are various potential sources for an eventual lack of measurement invariance. Such sources threatening the invariance of the scale may result, for example, from suboptimal translations, a different understanding of various question items, or cultural variations in response style. We present the results of the invariance test below.

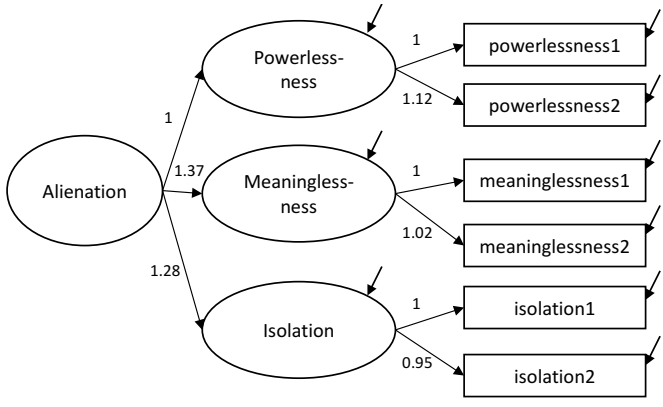

Figure 1 The second-order factor measurement model of alienation. The numbers are invariant unstandardized factor loadings as estimated in a second-order metric invariance model (corresponding to Model 3 in Table 3).

Table 2 Indicators of Alienation Used in the "Group-Focused Enmity" Survey

Second-order concept	First-order concept	Definition (Seeman, 1959)	Questionnaire items, each with four response options: 1 – "Strongly agree" 2 – "Somewhat agree" 3 – "Somewhat disagree" 4 – "Strongly disagree"
Alienation	Powerlessness	individual's sense of influence over socio-political events	1) Politicians do not care what people like me think 2) People like me do not have any say about what the government does
	Meaninglessness	when the individual is unclear on what s/he ought to believe – when the individual's standards for clarity in decision making are not met	1) Nowadays things are so confusing that you sometimes do not know where you stand 2) Nowadays things are so complex that you sometimes do not know what is going on
	Social Isolation	alienation from reigning goals and standards	1) Finding real friends is becoming more and more difficult nowadays 2) Relationships are getting more and more unstable

Method

To check whether we can compare the alienation scale across countries, we first specified a second-order confirmatory factor analysis model. It is depicted in Figure 1.

One loading of each first- and the second-order factor was fixed to 1 in order to identify the covariance structure part of the model (applying the marker item method). As we do not assess partial measurement invariance, the selection of marker indicators did not require any additional test of the adequacy of the chosen item. However, during the analysis we paid special attention to whether the modification indices suggest that the marker item's parameters are not equal across groups. As markers we selected the indicator "Politicians do not care what people like me think" for the powerlessness factor, the indicator "Nowadays things are so confusing that you sometimes do not know where you stand" for the meaninglessness factor, and the indicator "Finding real friends is becoming more and more dif-

ficult nowadays" for the social isolation factor. For the second-order factor, powerlessness was chosen to be the marker of the alienation factor because this first-order factor was treated as the very core of alienation in a number of studies (e.g., Geis & Ross, 1998; Neal & Seeman, 1964). Of all subdimensions, this one has been the most extensively studied (Seeman, 1975, p. 94). Moreover, Seeman (1959, p. 784) linked this concept to the original formulation of the alienation concept by Marx. Since our indicators had only four response options, the parameters were estimated using the maximum likelihood robust (MLR) estimator. In order to simplify the description, we treated these indicators as continuous.[6] All the models were tested using the software Mplus 7.3 (Muthén & Muthén, 1998-2016). The syntax codes are provided in Appendix D.

We began by fitting the CFA model in each country separately (not reported).[7] The model demonstrated an acceptable fit in all countries with the exception of Portugal. Consequently, we decided to exclude Portugal from further analysis. We checked invariance in five steps (as described in previous sections) according to the constraints listed in Table 1.

Results

Table 3 displays the fit measures of the five models we tested. Model 1, which included no cross-groups constraints, displayed a very good fit. Thus, we could conclude that each construct was measured by the same items in each of the countries included in the analysis. Also Model 2, which tested for metric invariance of the first-order factors, demonstrates a good fit. The χ^2 difference test suggested that there is no significant deterioration in the model fit compared to Model 1. In addition, the difference in CFI did not exceed 0.01. This indicates that the first-order factor loadings could be considered invariant across countries. Similarly, also Model 3, where we tested for metric invariance of the second-order factor, demonstrated a good fit. A comparison with Model 2 revealed no significant deterioration in the χ^2 value or in the CFI value. Therefore, we could conclude that the second-order factor loadings are invariant across countries as well. This finding implies the equal meaning of alienation across countries.

6 An examination of the item distributions did not detect any severe nonnormalities. In order to check the robustness of the results, we reanalyzed the model while accounting for the ordinal nature of the observed items using the WLSMV estimator in Mplus (see, e.g., Davidov et al., 2011). The model was identified using the constraints suggested by Millsap and Yun-Tein (2004), the second-order scalar invariance model was identified by constraining the latent intercepts to 0 in all groups and the second-order factor's mean to 0 in one group. The model fit indices are listed in Appendix B and demonstrate that our conclusions remain essentially the same.

7 The output may be obtained from the first author upon request.

Models 4 and 5 tested for full scalar invariance of the first- and second-order factors in the model. Imposed scalar invariance of the first-order factors in Model 4 showed a substantial deterioration in model fit both in terms of the χ^2 and the CFI. This finding implies that there is no first-order scalar invariance across all countries and, consequently, no second-order scalar invariance. However, for illustrative purposes, we also fitted a model testing for scalar invariance of the second-order factor in Model 5. As expected, this model showed a poor fit to the data. Thus, the best model in this sequence was Model 3, which demonstrated both first- and second-order metric invariance. Supporting these conclusions, the SABIC displayed the smallest value in this model as well.

As scalar invariance was not evidenced for both the first- and second-order factors in the model, means of the three first-order factors of alienation as well as the mean of the second-order factor of alienation may not be compared with confidence across countries. Since we only had two items measuring each first-order factor, and as partial scalar invariance requires that at least two items per factor display equal factor loadings and intercepts, it was not possible for us to test for partial scalar invariance.

Lack of evidence of scalar measurement invariance does not necessarily imply that no comparisons can be performed. It could well be the case that although the first- and second-order factors of alienation may not be comparable across all eight countries, there are pairs or triads of countries where they are comparable and where scalar invariance can be supported by the data. For example, we found full scalar invariance of this model between Italy and Germany (the fit indices are listed in Appendix C). The mean alienation in Italy was 0.344 and significant, whereas in Germany the mean was fixed to 0. Thus, the level of alienation was significantly higher in Italy than in Germany. Furthermore, we found empirical support for partial scalar invariance across Poland and France. In this model, the latent intercept of the first-order factor of meaninglessness was freed, whereas the intercepts of the first-order factors powerlessness and isolation were constrained to equality. The mean alienation in Poland was 0.471 and significant, whereas in France the mean was fixed to 0. In line with our expectations, the level of alienation is significantly higher in Poland than in France. Likewise, we found partial scalar invariance for Germany and the United Kingdom. In the model for these two countries we had to relax intercepts of the observed indicator of the first-order factor isolation, as well as the latent intercept of isolation itself. In the United Kingdom the latent mean of alienation was fixed to 0, whereas in Germany it was estimated as -0.160 and highly significant, indicating that the level of alienation was higher in the United Kingdom. Researchers interested in studying specific countries in these data would need to conduct the analysis we presented for these particular countries to determine whether they exhibit full or partial invariance.

Table 3 Results of Invariance Tests of a Second-Order Factor Model of Alienation

	$\chi^2(df)$	Scaled χ^2 difference	CFI	CFI difference	RMSEA	SRMR	SABIC
1) Configural invariance	49.5 (42)		0.998		0.019 [a]	0.014	51295
2) Metric invariance of the first-order factors	71 (60)	21.6	0.997	0.001	0.019 [a]	0.025	51231
3) Metric invariance of the first and second-order factors	79.8 (72)	8.6	0.997	0.001	0.015 [a]	0.029	51181
4) Scalar invariance of the first-order factors	417.2 (90)*	337.4*	0.917	0.080	0.085	0.063	51483
5) Scalar invariance of the first- and second-order factors	691.9 (102)*	274.2*	0.850	0.063	0.107	0.094	51740

Note. df – degrees of freedom; scaled $\chi 2$ difference is a difference between -2log-likelihood corrected with a scaling factor applied with maximum likelihood robust estimator; CFI – comparative fit index; delta CFI – difference in CFI from the previous model in the sequence; RMSEA – root mean square error of approximation, SABIC – sample-adjusted Bayesian information criterion, SRMR – standardized root mean square residual.

* significant at $p < 0.01$.

a – RMSEA is equal or lower than 0.05 at $p < 0.05$ level of significance.

Summary and Conclusions

Measurement invariance is a necessary condition to allow meaningful comparisons across groups. The last two decades have witnessed a significant increase in the number of cross-cultural studies which tested for measurement invariance across groups such as cultures, countries, or language groups (Davidov et al., 2014). MGCFA is currently one of the most common techniques used for assessing measurement invariance. However, higher-order factor modeling was only seldom discussed. In particular, the literature has provided only very general guidelines for testing measurement invariance of second-order factor models (and of higher-order factors in general). This is unfortunate, because measurement invariance is also

a prerequisite for conducting meaningful comparative research when second- (or higher-) order factors are included in a study. In an attempt to fill this gap, the current paper first presents a nontechnical explanation of the constraints required for the identification of models and the different steps that are taken when testing for measurement invariance of second-order factors in a multiple-group model. Second, it provides a practical application of how to test for measurement invariance of a second-order factor using data drawn from eight European countries. It measures the second-order concept of alienation with its three first-order dimensions: powerlessness, meaninglessness, and social isolation.

The empirical example was performed using the concept of alienation as a second-order factor, where meaninglessness, powerlessness, and isolation served as first-order factors, each measured by two indicators. We found support for first- and second-order metric invariance among seven countries (excluding Portugal), but no support for scalar invariance across countries. Does it imply that alienation may not be compared across all countries? Strictly speaking, at least partial scalar invariance for the first- and second-order factors is necessary to guarantee that mean comparisons of alienation across countries are meaningful. However, we suggest that differences in the structural parameters for the second-order factors (e.g., differences in the intercepts of the first-order factors across countries) may reveal that the concept of alienation bears somewhat different connotations and content across countries. This could be a useful starting point for substantive researchers to examine reasons for the revealed parameter differences across countries.

The criteria described in this paper to test for measurement invariance require exact equality of factor loadings and intercepts. In recent times, however, this approach has often been regarded as too strict. For this reason, novel and more lenient forms of measurement invariance methods such as approximate Bayesian invariance (Muthén & Asparouhov, 2013) or alignment (Asparouhov & Muthén, 2014) are gaining popularity. Although these new methods are very promising, they are beyond the scope of the current paper. These newer procedures may suggest that scales are (approximately and sufficiently) invariant even when exact measurement invariance tests fail to do so. Such approximate invariance tests can also take into account parameters differences across countries in a more flexible way than our approach does. As we are not aware of any studies that have applied these procedures on second- or higher-order factors, a task for future studies is to do so and to provide illustrations of how to assess approximate measurement invariance for higher-order factors.

The study has several limitations related both to our measurements and the criteria used to assess measurement invariance. Measures were only available for three of the five subdimensions of our second-order factor of alienation. Thus, we could test its measurement invariance properties while only reflecting a part of its subdimensions. In addition, each first-order factor was measured by only two

items. Thus, it was not possible for us to test whether partial (rather than full) scalar invariance was given in the data for the first-order factors. A test of partial invariance requires having at least three indicators to measure each first-order factor. However, the data we used also offered several advantages. In particular, the data represent a realistic and common situation in survey research in which we have only two items to measure each latent variable (see, e.g., the case of the value measurements in the European Social Survey). Second, the simplicity of the data allows for a clearer illustration of the procedure. Third, the illustration presented here uses data on an important concept in sociological and social psychological literature. Fourth, the data allow for testing a second-order factor across a large number of countries. An additional limitation we would like to acknowledge is that it is not clear whether the criteria we used to determine whether measurement invariance models are supported by the data, such as exploring differences in CFI across models (Chen, 2007; see also Cheung & Rensvold, 2002), apply also for models testing for measurement invariance of *second-order* factors. These criteria were developed originally for models with first-order factors. Future simulation studies may assess whether these criteria also apply for the test of measurement invariance of second-order factors. In spite of these limitations, we believe that testing for measurement invariance of a second-order factor is essential when using data from multiple samples and comparing these latent variables across countries. We hope that the nontechnical presentation of this method reported in this article will help researchers in their endeavor to study second- (or higher-) order factors from a cross-cultural perspective.

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243-277). Mahwah, NJ: Lawrence Erlbaum Associates.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling, 21*(4), 495-508. doi:10.1080/10705511.2014.919210

Boorsbom, D., Mellenbergh, J. G., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203-219. doi:10.1037/0033-295X.110.2.203

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466. doi:10.1037/0033-2909.105.3.456

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling*, 14(3), 464-504. doi:10.1080/10705510701301834

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*(3), 471-492. doi:10.1207/s15328007sem1203_7

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*(2), 189-225. doi:10.1207/s15327906mbr4102_5

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. doi:10.1207/S15328007SEM0902_5

Cieciuch, J., Davidov, E., Vecchione, M., & Schwartz, S. H. (2014). A hierarchical structure of basic human values in a third-order confirmatory factor analysis. *Swiss Journal of Psychology, 73*(3), 177-182. doi:10.1024/1421-0185/a000134

Costa Jr, P. T., & McCrae, R. R. (1990). Personality disorders and the five-factor model of personality. *Journal of Personality Disorders, 4*(4), 362-371. doi:10.1521/pedi.1990.4.4.362

Costner, H. L. (1969). Theory, deduction, and rules of correspondence. *American Journal of Sociology*, *75*(2), 245-263. doi:10.1086/224770

Davidov, E., Datler, G., Schmidt, P., & Schwartz, S. H. (2011). Testing the invariance of values in the Benelux countries with the European Social Survey: Accounting for ordinality. In E. Davidov, P. Schmidt, & J. Billiet. (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 149-171). New York: Routledge.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55-75. doi:10.1146/annurev-soc-071913-043137

Davidov, E., Schmidt, P., & Billiet, J. (2011). *Cross-cultural analysis: Methods and applications*. New York: Routledge.

De Beuckelaer, A., & Swinnen, G. (2011). Biased latent variable mean comparisons due to measurement noninvariance: A simulation study. In E. Davidov, P. Schmidt, & J. Billiet. (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 117-148). New York: Routledge.

Dean, D. (1961). Alienation: Its meaning and measurement. *American Sociological Review, 26*(5), 753-758. doi:10.2307/2090204

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149. doi:10.1177/0748175610373459

Geis, K. J., & Ross, C. E. (1998). A new look at urban alienation: The effect of neighborhood disorder on perceived powerlessness. *Social Psychology Quarterly*, *61*(3), 232-246. doi:10.2307/2787110

Hempel, G. C. (1973). The meaning of theoretical terms: A critique of the standard empiricist construal. *Studies of Logic and the Foundation of Mathematics, 74*, 367-378. doi:10.1016/S0049-237X(09)70372-6

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117-144. doi:10.1080/03610739208253916

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1-55. doi:10.1080/10705519909540118

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*, 131-152. doi:10.1177/0022002185016002001

Huschka, D., & Mau, S. (2006). Social anomie and racial segregation in South Africa. *Social Indicators Research, 76*(3), 467-498. doi:10.1007/s11205-005-2903-x

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Greenwood Publishing Group.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102-138). New York: Guilford Press.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling, 16*(4), 642-657. doi:10.1080/10705510903206014

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409-426. doi:10.1007/bf02291366

Jung, E., & Yoon, M. (2017). Two-step approach to partial factorial invariance: Selecting a reference variable and identifying the source of noninvariance. *Structural Equation Modeling*, *24*(1), 65-79. doi:10.1080/10705511.2016.1251845

Kish, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, *44*(247), 380-387. doi:10.1080/01621459.1949.10483314

Krebs, D., & Schuessler, K. F. (1989). Life Feeling Scales for use in German and American samples. *Social Indicator Research, 21*(2), 113-131. Retrieved from http://www.jstor.org/stable/27520757

Küpper, B., Wolf, C., & Zick, A. (2010). Social status and anti-immigrant attitudes in Europe: An examination from the perspective of social dominance theory. *International Journal of Conflict and Violence*, *4*(2), 205-219. doi:10.4119/UNIBI/ijcv.85

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*(1), 59-72. doi:10.1207/s15328007sem1301_3

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*(2), 57-64. doi:10.1027/1614-2241.2.2.57

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320-341. doi:10.1207/s15328007sem1103_2

McClosky, H., & Schaar, J. H. (1965). Psychological dimensions of anomy. *American Sociological Review*, *30*(1), 14-40. doi:10.2307/2091771

McCutcheon, A. L. (1987). *Latent class analysis* (Sage University papers series on Quantitative Applications in the Social Sciences, No. 07-064). Newbury Park, CA: Sage.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543. doi:10.1007/bf02294825

Middleton, R. (1963). Alienation, race, and education. *American Sociological Review*, *28*(6), 973-977. Retrieved from http://www.jstor.org/stable/2090316

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39(3), 479-515. doi:10.1207/S15327906MBR3903_4

Muthén, B., & Asparouhov, T. (2013). *BSEM measurement invariance analysis.* Mplus Web Notes, 17. Retrieved from https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, L. K., & Muthén, B. O. (1998-2016). *Mplus user's guide. Sixth edition*. Los Angeles, CA: Muthén & Muthén.

Neal, A. G., & Seeman, M. (1964). Organizations and powerlessness: A test of the mediation hypothesis. *American Sociological Review*, *29*(2), 216-226. Retrieved from http://www.jstor.org/stable/2092124

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*(1), 51-67. doi:10.1207/s15327906mbr2301_3

Robinson, J. (1973). Alienation and anomie. In J. P. Robinson & P. R. Shaver (Eds.), *Measures of social psychological attitudes* (Rev. ed., pp. 245-294). Ann Arbor, MI: Institute for Social Research.

Schmidt, P. (1990). Measurement of powerlessness and alienation: Conceptual analysis, dimensionality and covariance structure analysis. In J. J. Hox & J. de Jong-Gierveld (Eds.), *Operationalization and research strategy* (pp. 103-122). Amsterdam: Swets & Zeitlinger.

Schwartz, S. H., & Bilsky, W. (1990). Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications. *Journal of Personality and Social Psychology*, *58*(5), 878-891. doi:10.1037/0022-3514.58.5.878

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., Dirilen-Gumus, O., & Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology, 103*, 663-688. doi:10.1037/a0029393

Seeman, M. (1959). On the meaning of alienation. *American Sociological Review*, *24*(6), 783-791. Retrieved from http://www.jstor.org/stable/2088565

Seeman, M. (1975). Alienation studies. *Annual Review of Sociology, 1*(1), 91-123. doi:10.1146/annurev.so.01.080175.000515

Seeman, M. (1983). Alienation motifs in contemporary theorizing: The hidden continuity in the classic themes. *Social Psychology Quarterly, 46*(3), 171-184. Retrieved from http://www.jstor.org/stable/3033789

Seeman, M. (1991). Alienation and anomie. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 291-372). San Diego, CA: Academic Press.

Steinmetz, H. (2011). Estimation and comparison of latent means across cultures. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 85-116). New York: Routledge.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz S. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality and Quantity, 43*, 599-616. Retrieved from https://doi.org/10.1007/s11135-007-9143-x

Strasheim, A. (2011). Testing the invariance of second-order confirmatory factor analysis models that include means and intercepts. *Management Dynamics: Journal of the Sou-*

thern African Institute for Management Scientists, 20(4), 38-75. Retrieved from http://hdl.handle.net/10520/EJC119148

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486-492. doi:10.1080/17405629.2012.686740

Van de Schoot, R., Kluytmans A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4,* 770. 10.3389/fpsychg.2013.00770

Van de Vijver, F. J., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Newbury Park, CA: Sage.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70. doi:10.1177/109442810031002

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209-231). New York: Guilford Press.

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, 81(4), 1014-1045. doi:10.1007/s11336-016-9506-0

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*(2), 113-128. doi:10.1007/BF02294531

Zick, A., Küpper, B., & Hövermann, A. (2011). *Intolerance, prejudices and discrimination – A European report*. Berlin: Friedrich-Ebert Foundation.

Appendix A

Sample Characteristics

Country	Response rate, %	Sample size	Percentage females	Average age
France	10.2	531	53.4	46.0
Great Britain	24.6	519	50.6	46.8
Germany	33.0	495	50.2	47.9
Hungary	8.8	477	50.9	46.9
Italy	4.5	499	50.6	49.9
Netherlands	11.8	513	49.4	46.9
Portugal	7.3	483	52.9	45.4
Poland	15.5	501	52.3	43.1

Appendix B

Fit Indices of a Measurement Invariance Test of the Second-Order Factor of Alienation while Accounting for Ordinality of the Items (Using the WLSMV Estimator)

	$\chi^2(df)$	χ^2 difference	CFI	CFI difference	RMSEA	RMSEA upper boundary
1) Configural invariance	63.1 (42)		0.999		0.032	0.047
2) Metric invariance of the first-order factors	138.1 (60)	62.4*	0.993	0.007	0.051	0.062
3) Metric invariance of the first- and second-order factors	149.1 (72)	19.0	0.994	0.001	0.046	0.057
4) Scalar invariance of the first-order factors	634.5 (126)	519.1*	0.978	0.016	0.090	0.097
5) Scalar invariance of the first- and second-order factors	1122.4 (138)	309.2*	0.949	0.031	0.119	0.126

Note. * significant at $p < 0.01$ as estimated by *DIFFTEST* procedure in Mplus.

Appendix C

Fit Indices of the Second-Order Factor Models of Alienation in Italy and Germany

	$\chi^2(df)$	Scaled χ^2 difference	CFI	CFI difference	RMSEA	SRMR	SABIC
1) Configural invariance	15.0 (12)		0.997		0.023	0.015	14370
2) Metric invariance of the first-order factors	21.7 (15)	6.62	0.996	0.001	0.024	0.027	14367
3) Metric invariance of the first- and second-order factors	22.0 (17)	0.37	0.994	0.002	0.030	0.027	14360
4) Scalar invariance of the first-order factors	29.6 (20)	7.59	0.992	0.002	0.031	0.031	14357
5) Scalar invariance of the first- and second-order factors	35.2 (22)	5.62*	0.989	0.003	0.035	0.034	14356

* significant at $p < 0.01$.

Appendix D

Mplus Codes

1. Configural invariance model

```
DATA:
  FILE IS alienation7countries.dat;

VARIABLE:
  NAMES ARE country power1 power2 meaning1 meaning2 isolat1 isolat2;
  MISSING IS power1 power2 meaning1 meaning2 isolat1 isolat2 (5);
  GROUPING IS country (1=GB 2=GE 3=HU 4=IT 5=NE 7=PL 8=FR);

ANALYSIS:
  ESTIMATOR = MLR;

MODEL:
  POWER   BY power1@1 power2;
  ISOLAT  BY isolat1@1 isolat2;
  MEANING BY meaning1@1 meaning2;
  ALIENAT BY POWER@1 ISOLAT MEANING;

MODEL GB: !This block is repeated for each country
  POWER   BY power2;
  ISOLAT  BY isolat2;
  MEANING BY meaning2;

  [power1@0 power2 isolat1@0 isolat2 meaning1@0 meaning2];

  ALIENAT BY ISOLAT MEANING;
  [POWER ISOLAT MEANING];
  [ALIENAT@0];
```

2. Metric invariance of the first-order factors. Data, variable, and analysis blocks are the same as in the configural model). Hereafter, the additions to the code of the preceding model are in bold.

```
MODEL GB: !This block is repeated for each country
  POWER   BY power2  (load1);
  ISOLAT  BY isolat2 (load2);
  MEANING BY meaning2(load3);

  [power1@0 power2 isolat1@0 isolat2 meaning1@0 meaning2];

  ALIENAT BY ISOLAT MEANING;
  [POWER ISOLAT MEANING];
  [ALIENAT@0];
```

3. Metric invariance of the first- and second-order factors

```
MODEL GB: !This block is repeated for each country
  POWER   BY power2  (load1);
  ISOLAT  BY isolat2 (load2);
  MEANING BY meaning2(load3);

  [power1@0 power2 isolat1@0 isolat2 meaning1@0 meaning2];

  ALIENAT BY ISOLAT MEANING (load4 load5);
  [POWER ISOLAT MEANING];
  [ALIENAT@0];
```

4. Scalar invariance of the first-order factors

```
MODEL GB: !This block is repeated for each country
  POWER   BY power2  (load1);
  ISOLAT  BY isolat2 (load2);
  MEANING BY meaning2(load3);

  [power1@0 power2 isolat1@0
  isolat2 meaning1@0 meaning2] (intcpt1-intcpt6);

  ALIENAT BY ISOLAT MEANING (load4 load5);
  [POWER ISOLAT MEANING];
  [ALIENAT@0];
```

5. Scalar invariance of the first- and second-order factors.

```
MODEL GB: !This block is repeated for each country
  POWER   BY power2  (load1);
  ISOLAT  BY isolat2 (load2);
  MEANING BY meaning2(load3);

  [power1@0 power2 isolat1@0
  isolat2 meaning1@0 meaning2](intcpt1-intcpt6);

  ALIENAT BY ISOLAT MEANING (load4 load5);
  [POWER ISOLAT MEANING](intcpt7-intcpt9);
  [ALIENAT@0]; ! This line should be [ALIENAT*] in all the other
groups, i.e. latent mean is freely estimated except for one group.
```

Appendix E

Replication data. Variances and covariance matrices and means for the manifest variables in each country.

	POWER1	POWER2	MEANING1	MEANING2	ISOLAT1	ISOLAT2
Great Britain						
POWER1	0.88					
POWER2	0.58	0.99				
MEANING1	0.25	0.31	0.92			
MEANING2	0.19	0.25	0.62	0.87		
ISOLAT1	0.22	0.24	0.27	0.27	1.12	
ISOLAT2	0.20	0.18	0.29	0.22	0.35	0.81
Means	2.86	2.84	2.90	2.95	2.26	2.91
Germany						
POWER1	0.91					
POWER2	0.54	0.97				
MEANING1	0.30	0.29	0.86			
MEANING2	0.29	0.31	0.67	0.88		
ISOLAT1	0.30	0.30	0.37	0.43	1.01	
ISOLAT2	0.20	0.24	0.27	0.31	0.40	0.75
Means	2.87	2.77	2.60	2.64	2.67	2.84
Hungary						
POWER1	0.80					
POWER2	0.25	1.31				
MEANING1	0.24	0.31	1.04			
MEANING2	0.28	0.25	0.65	0.94		
ISOLAT1	0.21	0.16	0.31	0.29	0.98	
ISOLAT2	0.19	0.16	0.37	0.34	0.55	0.91
Means	3.27	2.43	2.97	3.08	3.17	3.21
Italy						
POWER1	0.72					
POWER2	0.36	0.77				
MEANING1	0.24	0.21	1.02			
MEANING2	0.20	0.22	0.63	0.83		
ISOLAT1	0.13	0.26	0.28	0.25	1.00	
ISOLAT2	0.18	0.24	0.26	0.23	0.54	0.79
Means	3.19	3.32	3.06	3.13	3.08	3.13

	POWER1	POWER2	MEANING1	MEANING2	ISOLAT1	ISOLAT2
Netherlands						
POWER1	0.84					
POWER2	0.57	0.88				
MEANING1	0.15	0.17	0.78			
MEANING2	0.15	0.20	0.50	0.74		
ISOLAT1	0.17	0.21	0.25	0.20	0.88	
ISOLAT2	0.18	0.21	0.20	0.22	0.42	0.79
Means	2.19	2.36	2.67	2.71	2.14	2.62
Poland						
POWER1	0.68					
POWER2	0.38	0.80				
MEANING1	0.12	0.15	0.64			
MEANING2	0.15	0.14	0.46	0.70		
ISOLAT1	0.12	0.17	0.21	0.25	0.83	
ISOLAT2	0.15	0.14	0.18	0.19	0.29	0.55
Means	3.36	3.31	3.28	3.13	3.12	3.34
France						
POWER1	0.96					
POWER2	0.58	1.18				
MEANING1	0.26	0.27	0.79			
MEANING2	0.30	0.36	0.54	0.79		
ISOLAT1	0.28	0.29	0.32	0.34	1.21	
ISOLAT2	0.31	0.32	0.34	0.33	0.86	1.15
Means	2.91	2.69	3.18	3.08	2.52	2.77

methods, data, analyses | Vol. 12(1), 2018, pp. 77-104 DOI: 10.12758/mda.2017.09

Using Alignment Optimization to Test the Measurement Invariance of Gender Role Attitudes in 59 Countries

Vera Lomazzi
GESIS - Leibniz Institute for the Social Sciences

Abstract

Several repeated cross-national surveys include measurements of attitudes toward gender roles to investigate individuals' beliefs regarding the appropriateness of men and women's roles in a particular context. When used to compare attitudes across countries, these measurements reveal critical factors that could cause a lack of equivalence between different cultural contexts, and that could therefore produce misleading results. Nevertheless, the use of such measures to compare country means without assessing measurement equivalence is common. It should also be considered that the assessment of equivalence within a large-scale sample from cross-sectional surveys through multigroup confirmatory factor analysis (MGCFA) often fails because of the strict requirements necessary.

The current article is used to assess the measurement equivalence of the gender role attitudes scale included in the last wave of the World Values Survey in 59 countries, with the main goal of identifying the most invariant model for the largest number of groups. The study involved comparing two methods belonging to the frequentist approach: MGCFA and the frequentist alignment procedure, a highly novel and promising method that is still rarely used. Using the first technique, partial scalar invariance was achieved for 27 countries. By employing the frequentist alignment optimization, an acceptable degree of non-invariance was achieved for 35 countries. Thus, the study confirmed the frequentist alignment procedure as a viable alternative to the MGCFA.

Keywords: Alignment; measurement invariance; measurement equivalence; World Values Survey; gender role attitudes; multigroup confirmatory factor analysis

Introduction

Scholars have been well aware of the relevance of the comparative perspective since the dawn of sociology. From Durkheim and Weber onward, the comparative approach has been adopted to highlight differences and similarities among different groups in an attempt to make theoretical generalizations. This approach is grounded in the basic assumption of comparability; however, are we really comparing the same thing across the different groups?

In the field of survey research, this concern is intertwined with the issue of measurement equivalence and the methodological approaches used to test for it. According to Horn and McArdle (1992, p. 117), the question of measurement invariance is one of "whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute." If measurement invariance is lacking, results can be misinterpreted and conclusions led by "methodological artefacts" (Moors, 2004).

In recent decades, the development of several cross-cultural and repeated survey programs has increased the possibilities for comparative research, both across cultural groups and over time. The efforts made by these programs to guarantee the quality of the data collected lead to the provision of more reliable data, but numerous issues can arise that result in the lack of effective equivalence. In addition to the common causes of non-invariance, such as differences in modes of data collection, sampling, and translation issues (van de Vijver & Tanzer, 2004), cultural biases could arise from the different interpretations of the questions; furthermore, social desirability and acquiescence can also differ by context (Heath, Martin, & Spreckelsen, 2009). The risk of comparing "apples and oranges," as raised by Stegmueller (2011), is therefore always in play. The scientific discourse in this field has recently been reinvigorated by two emerging debates, one questioning formative versus reflexive approaches to the study of latent concepts, and the other addressing the exact versus approximate approaches to the concept of equivalence itself, with the consequential development of new techniques to assess invariance.

Scholars such as Welzel and Inglehart (Inglehart & Welzel, 2005; Welzel, 2013; Welzel & Inglehart, 2016) have assumed a formative approach to the cross-cultural study of values. Against the "dimensional logic" commonly adopted by the reflexive approach, which considers item responses as reflections of latent concepts, they proposed a "combinatory logic." In other words, their measures of values are defined following a theoretical perspective, as they select items to build composite indexes. Nevertheless, these authors have in their previous studies used methods

Direct correspondence to
Vera Lomazzi, GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
E-mail: vera.lomazzi@gesis.org

that are only applicable for reflective indicators on the same indicators that they claim to be formative, and thus have made their argument less convincing. An example of this can be seen in the paper by Inglehart and Baker (2000) in which the authors aimed to test the postmaterialism theory in 43 societies. They identified 10 items selected from the World Values Survey carried out in 1990–91 and 1995–98 that tap the "Traditional vs. Secular-rational Values" and the "Survival vs. Self-expression Values", following their combinatory logic. However, to demonstrate that these two dimensions of cross-cultural variations exist both at national and individual levels, they then used a factor model, which is a technique for dealing with reflective indicators.

In addition, as pointed out by van Vlimmeren, Moors, and Gelissen (2016), the formative approach emphasizes the researcher's point of view; thus, the index could measure the concept as it is framed in the social researcher's mind, neglecting what is going on in the minds of respondents and the fact that the meaning given to that item, or the way of responding, can be culturally dependent. Welzel's approach has also been criticized because it underestimates the problem of cross-cultural equivalence and measurement errors (Alemán & Woods, 2015; van Deth, 2014; van Vlimmeren et al., 2016).

Scholars who refer to dimensional logic have strongly argued for the importance of equivalence in comparative studies. Alemán and Woods (2016) widely demonstrated that the postmaterialism and emancipative measures built through the formative approach are not equivalent. In their response, Welzel and Inglehart (2016) expressed the idea that measurement invariance is overrated and is not necessary when adopting a combinatory logic; instead, convergence with external criteria is sufficient to validate the measure and use it at the aggregate level.

Meanwhile, novel approaches to address measurement invariance have been emerging. Contrasting with the exact approach, which requires "exact equivalence" between parameters, the current development of the assessment of measurement invariance refers to the concept of "approximate equivalence," which includes cultural variability and uncertainty in the assessment (Muthén & Asparouhov, 2013; van de Schoot et al., 2013). In the frame of this debate, the alignment method (Asparouhov & Muthén, 2014) has been proposed to conveniently compare means, introducing the idea that a certain amount of non-invariance is acceptable. This procedure, which can be employed in both the exact and the approximate approaches to equivalence, appears to be particularly useful when handling data from a large number of groups (Kline, 2015; Muthén & Asparouhov, 2014). Nevertheless, only a few studies have already applied this new approach to substantive research and, at the same time, the evaluation of the measurement invariance of gender role attitudes remains rare, even if these measures are often used to compare support for gender equality across countries.

The present study, which adopted the reflective approach, had a two-fold goal. The first was to assess the measurement invariance of gender role attitudes by identifying the most invariant model across the largest group of countries among those available in the sixth wave of the World Values Survey (WVS). The second was to explore two different methods to assess equivalence, both belonging to the frequentist approach; in addition to MGCFA, the new frequentist alignment optimization was also adopted, and the results then compared.

Approaches to Measurement Invariance

Among the methods often employed to assess measurement invariance, including latent class modeling (Kankaraš & Moors, 2009) and item response theory (Millsap, 2010), MGCFA has been the most commonly used (Davidov et al., 2015). These methods refer to the traditional approach to measurement invariance, which has its roots in the concept of "exact equivalence." In other words, the test of general theories and the comparison between different groups will be successful if the instrument used to compare them is exactly the same.

Previous studies have referred to three levels of measurement invariance: configural, metric, and scalar (Steenkamp & Baumgartner, 1998). The first of these refers to the fact that the construct responds to the same configuration in all groups; in other words, the same pattern of factor loading is shown across the groups. Metric invariance requires that the unit of measurement is the same, so that the factor loadings are constrained to be equal across the groups. The third level of invariance is the most demanding, as scalar invariance requires equality in factor loadings and indicator intercepts. Comparing covariances and unstandardized regression coefficients across the groups is also possible when metric invariance is reached, but only by achieving scalar invariance can the latent means be compared (Davidov, 2010; Steenkamp & Baumgartner, 1998). However, Byrne et al. (1989) and Steenkamp and Baumgartner (1998) argued that partial invariance is also an acceptable condition for comparing means. In this case, at least two items with equal parameters (factor loadings for partial metric invariance, and factor loading and intercepts for partial scalar invariance) must be identified.

Although the concept of invariance is fundamental in allowing meaningful mean comparisons, some studies have recently claimed that the classical "exact" approach to equivalence presents some problems (Asparouhov & Muthén, 2014; Davidov et al., 2015; Muthén & Asparouhov, 2013; Van De Schoot et al., 2013). When addressing a large number of groups, which is often the case in large-scale cross-national surveys, the traditional approach is too strict, rejecting models that are practically comparable across groups (for example, where the countries' mean ranking is not biased although the parameters are not exactly equal) and hard to

fulfill. It is often impossible to achieve full invariance since the possible violations in terms of equivalence increase as the number of groups is increased (Davidov, Meuleman, Billiet, & Schmidt, 2008; Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014). Researchers must employ a lengthy procedure to identify an acceptable partially invariant model, which generally requires numerous large modification indexes; however, these modifications can lead to the risk of producing an inappropriate model because of "the scalar model being far from the true model," as pointed out by Asparouhov and Muthén (2014, p. 495). Marsh et al. (2017, pp. 10–12) clearly explained this issue, which concerns the problems caused by the stepwise approach that leads to achieving partial invariance. The main argument is that the achievement of a good fit by freeing parameters does not guarantee that means are unbiased. In addition, because of the multicollinearity in the modification indices, the selection of the parameters to be freed risks being arbitrary and thus overlooking other potentially better models.

To avoid these risks, another pragmatic solution is to reduce the number of groups compared, but this also reduces the possibility of substantive analyses, with the consequential risks of comparing groups that tend to be culturally more similar and discarding groups that may be of real interest to the scholar.

To express this as well as van de Schoot et al. (2013), researchers find themselves caught between the two "monsters" of Scylla and Charybdis. Scylla, the six-headed monster, frightens scholars by imposing a model that, to achieve measurement invariance, poorly fits the actual data; Charybdis scares them with a model that, while fitting the data, is not invariant. Nowadays, the concept of "approximate equivalence" introduced by Muthén and Asparouhov (2012, 2013), appears to be the most feasible way of navigating between the two mythological monsters.

The two approaches rely on different assumptions. In the exact approach, the differences between factor loadings/intercepts among the groups are zero: they are exactly equal among the groups. In contrast, approximate equivalence considers that loadings/intercepts do not have to be identical among groups that are culturally different. This means that, even if the mean of the loadings/intercepts variations is zero, some slight differences are permitted. The recently developed alignment optimization can be employed in both the approximate/Bayesian and the exact/frequentist framework. In the latter case, its use could be particularly fitting for those who prefer to stick to the frequentist approach but skip the aforementioned problems caused by the stepwise process employed to achieve partial invariance.

While the application of different techniques in the Bayesian framework has attracted scholars' attention (Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014; Davidov et al., 2015; van de Schoot et al., 2013; Zercher, Schmidt, Cieciuch, & Davidov, 2015), the use of the frequentist alignment optimization (Asparouhov & Muthén, 2014) remains rarely applied. Therefore, the current study

aims to contribute to the exploration of this new method to assess measurement equivalence.

Alignment Optimization

Developed by Asparouhov and Muthén (2014) as an alternative to MGCFA, this method estimates the factor means without constraining loadings and equal intercepts across groups, and it discovers the most optimal measurement invariant pattern.

Different from the MGCFA, which assumes measurement invariance, the basic assumption of the alignment is that the number of non-invariant parameters and the degree of non-invariance can be kept to a minimum. This allows for finding an invariant pattern across the groups, and for estimating factor means and variances while considering the real differences in loadings and intercepts among groups. As a complementary output, the alignment procedure provides elements to assess the degree of non-invariance, which is helpful in evaluating whether to trust and accept the alignment results.

The frequentist alignment optimization technique begins by adopting the maximum likelihood (ML) method to estimate the configural model, where parameters do not all have to be equal, with factor means fixed at zero and factor variances fixed at one. This is model zero, the best-fitting model possible among the groups included in the analysis, without any restrictions on the parameters. After the optimization procedure, which involves applying a simplicity function that essentially works as the rotation criteria for the exploratory factor analysis (Asparouhov & Muthén, 2014, pp. 496–498), the final model retains the same fit as the configural model (model zero) but minimizes the amount of non-invariance.

Asparouhov and Muthén (2014; Muthén & Asparouhov, 2014) corroborated the validity of these techniques by conducting several Monte Carlo simulations. Monte Carlo simulation studies are generally employed to investigate the performance of statistical estimations in different conditions through the generation of multiple simulated samples of data from a defined population based on an assumed data-generating process (DGP) (Carsey & Harden, 2013). Asparouhov and Muthén (2014; Muthén & Asparouhov, 2014) used this feature to assess the performance of the alignment procedure in different settings. With regard to the amount of non-invariance that can be allowed without undermining the reliability of comparing the factor means, Asparouhov and Muthén (2014) stated that up to 20% of the parameters may be non-invariant for a researcher to be able to rely on the mean estimates. In further simulations, the authors (Muthén & Asparouhov, 2014, p. 3) raised the limit to 25%. They also recommended complementing the alignment measurement invariance assessment with Monte Carlo investigations when the level of non-invariance is higher.

The Measurement of Gender Role Attitudes in Comparative Research

The measurement of gender role attitudes appears to be particularly sensitive to construct bias, which occurs when "the construct measured is not identical across cultural groups" (van de Vijver & Tanzer, 2004, p. 120). In fact, different ways of defining gender roles are established across cultural contexts; institutional factors such as welfare regimes, religious traditions, or labor market dynamics have historically contributed to the development of different gender cultures across societies, prescribing gender roles accordingly (André, Gesthuizen, & Scheepers, 2013; Lomazzi, 2017a; Sjöberg, 2004). This is reflected not only in the shaping of gender beliefs, but also in the meaning given to the questions used to investigate these concepts (Braun, 1998, 2009), with the consequential result of a lack of equivalence between different cultural contexts, and therefore misleading results.

Irrespective of such a potential risk, the use of these measurements in comparative studies is relatively widespread. Only recent studies have introduced the evaluation of the quality of the measurement instruments in this field. Lomazzi (2017b) evaluated the cross-sectional reliability and stability of the configural structure of the gender role attitudes scale employed by the European Values Study across 26 countries, addressing caution in the use of the scale because not enough of it is tenable. Van Vlimmeren, Moors, and Gelissen (2016) recently analyzed family values and gender role items from the 2008 European Values Study, adopting the perspective of clusters of cultures to address the variation in the meaning given to items and in the way people who belong to different cultures answer the same questions. They clustered countries according to their similarity in covariances between items, and showed that such clusters are internally more invariant and then more comparable. Constantin and Voicu (2014) tested the invariance of the gender role scales included in the 2002 International Social Survey Programme (32 countries) and in the 2005 WVS (45 countries) using MGCFA. Their results showed that scalar invariance was not achieved in either case.

When comparing a large number of groups and, moreover, when the construct is particularly sensitive to situated social change, as in the case of gender beliefs (Braun, 1998, 2009; Constantin & Voicu, 2014; Lomazzi, 2017a), the traditional methods used to test invariance often fail (Asparouhov & Muthén, 2014; Davidov et al., 2015). Could a new method provide more encouraging results?

The Current Study

The aim in the present study was to assess the measurement invariance of the gender role attitudes scale employed by the last wave of the WVS, and to explore the limitations and potential of different methods in this assessment.

It has been suggested that the frequentist alignment method is highly convenient when analyzing several cultural groups (Kline, 2015; Muthén & Asparouhov, 2014). It also allows for overcoming the problems of the dubious model related to the achievement of partial invariance through MGCFA; therefore, in addition to the traditional MGCFA, its use appeared to be appropriate in the present study. Following a step-by-step procedure, the frequentist alignment optimization was employed to identify the best invariant model for as many groups as possible.

Methods

Data and Measurements

The study considered 59 of the 60 countries investigated by the sixth wave of the WVS (2015), giving a total sample size of 89,320 respondents (Argentina was excluded from the analyses because it had no valid case in one of the measures of interest). Table 1 shows each country's sample sizes and the country codes later used as references in the alignment output.

Table 1 Reference code and sample size by country

Code	Country	N
12	Algeria	1200
31	Azerbaijan	1002
36	Australia	1477
48	Bahrain	1200
51	Armenia	1100
76	Brazil	1486
112	Belarus	1535
152	Chile	1000
156	China	2300
158	Taiwan	1238
170	Colombia	1512
196	Cyprus	1000
218	Ecuador	1202
233	Estonia	1533
268	Georgia	1202
275	Palestine	1000

Code	Country	N
276	Germany	2046
288	Ghana	1552
344	Hong Kong	1000
356	India	5659
368	Iraq	1200
392	Japan	2443
398	Kazakhstan	1500
400	Jordan	1200
410	South Korea	1200
414	Kuwait	1303
417	Kyrgyzstan	1500
422	Lebanon	1200
434	Libya	2131
458	Malaysia	1300
484	Mexico	2000
504	Morocco	1200
528	Netherlands	1902
554	New Zealand	841
566	Nigeria	1759
586	Pakistan	1200
604	Peru	1210
608	Philippines	1200
616	Poland	966
634	Qatar	1060
642	Romania	1503
643	Russia	2500
646	Rwanda	1527
702	Singapore	1972
705	Slovenia	1069
710	South Africa	3531
716	Zimbabwe	1500
724	Spain	1189
752	Sweden	1206
764	Thailand	1200
780	Trinidad and Tobago	999
788	Tunisia	1205
792	Turkey	1605
804	Ukraine	1500
818	Egypt	1523
840	United States	2232
858	Uruguay	1000
860	Uzbekistan	1500
887	Yemen	1000
	Total	89320

Data: WVS, 2010-2014 (World Values Survey Association, 2015)

Gender role attitudes were measured through a battery of items, formulated as follows: 1) One of my main goals in life has been to make my parents proud (v49); 2) When a mother works for pay, the children suffer (v50); 3) On the whole, men make better political leaders than women (v51); 4) A university education is more important for a boy than for a girl (v52); 5) On the whole, men make better businesses executives than women (v53); and 6) Being a housewife is just as fulfilling as working for pay (v54). Responses to these statements were rated using scores ranging from 1, "Strongly agree," to 4, "Strongly disagree."

A preliminary exploratory factor analysis showed that the first item ("One of my main goals in life has been to make my parents proud") was far from belonging to the same latent concept of the scale (see Table A.1 in the Appendix). This was already imaginable from the content, as it related to feelings toward parents rather than to gender roles. Therefore, this item was not included in further analyses. The other five items were loaded on a unique factor, reflecting only one conceptual dimension.

Analysis Strategy

In order to achieve the two-fold goal of this study, the measurement equivalence was assessed in parallel, initially by performing MGCFA and then by employing the frequentist alignment method. In both cases, the Mplus 7.4 statistical modeling program (www.statmodel.com) was used and the same step-by-step procedure followed. Finally, the results obtained using the two techniques were discussed.

The criterion that guided the analytical strategy was the idea of finding a balance between the aim of including the biggest number of groups (ideally all those included in the survey) and the need for good enough coverage of the concept "attitudes towards gender roles" through the indicators included in the model.

In both procedures, the starting point was therefore the assessment of the 5-item model among all the available groups. Although prioritizing the ambitious aim of comparing as many countries as possible, when this first step did not allow for a reliable means comparison the second step was to identify the item that displayed the most non-invariant parameters and then exclude it from the measurement model. In this way, a 4-item model was identified and, again, the measurement equivalence was conducted across all the groups. A 3-item model was also considered, but because of several problems in the model identification, no further analyses were carried out. The strategy then included a third step, which aimed to identify an invariant measurement for a subset of groups.

In each of the three steps, the MGCFA was performed as follows. Initially, the model fit was assessed country-by-country, which eventually resulted in the exclusion of countries in which the fit was too poor. Then, full measurement invariance (all parameters constrained) was tested across the groups. When this was not

achieved, a close investigation of the modification indexes allowed identification of the most non-invariant parameters, which were gradually released to assess partial invariance. The measurement invariance was evaluated while considering the recommended cut-off criteria for the change in model fit: ΔCFI <0.01; ΔRMSEA <0.015; ΔSRMR <0.03 (Chen, 2007; Hu & Bentler, 1999). In the third main step, to reach an invariant measurement for a subset of groups, the most "problematic" groups (identified on the basis of the modification indices) were subsequently omitted.

Multigroup confirmatory factor analysis and the alignment method employ different computing procedures, which could result in different model fits, model identification, and, consequently, different subsets of groups. To assess the measurement equivalence using the frequentist alignment method, the analysis therefore began again using the original full sample.

The same procedure was applied at each of the three main steps; the alignment optimization was run using the ML estimator and the output was read to identify the amount of non-invariant parameters. Following the rule of thumb suggested by Muthén and Asparouhov (2014), a Monte Carlo investigation was performed to determine whether population values could be recovered via the alignment.

The Monte Carlo simulation was conducted using the parameters estimated by the alignment procedure as a data-generating population parameter values, defining a hypothetical sample of 1,500 units (the average sample size of the groups included in this study). This was performed both when the non-invariant rate was higher than 25%, as recommended by the developers of the alignment method (Muthén & Asparouhov, 2014), and also when this rate was lower, to validate this limit.

To select the item to be excluded using the measurement model (from step 1 to step 2) and the group to be dropped (from step 2 to step 3), the alignment optimization results were used as a diagnostic tool to identify the item (or group) that displayed the highest number of non-invariant parameters.

Results

The results are presented for both methods following the step-by-step procedure introduced earlier. For each model, the main results from the MGCFA and the alignment estimations are illustrated. For the latter, the full results and the Mplus excerpts (provided in the Appendix, Tables A.4 and A.5) are displayed only for the final models due to space limitations.

MGCFA Results

Table 2 summarizes the results from the first step using the traditional assessment of measurement equivalence of the 5-item model. For 2 of the 59 countries (Nigeria and Pakistan), the model fit was too poor, and these countries were excluded. The tests therefore refer to 57 countries. By releasing two factor loadings (v54, v52), partial metric invariance could be considered acceptable, even if the change in comparative fit index (CFI) was somewhat borderline (0.014). In order to test for partial scalar invariance, up to three intercepts were progressively released. However, this was not sufficient to establish partial scalar invariance; even if the changes in RMSEA and SRMS fitted the requirements, the change in CFI was higher than 0.01 (0.031). Moreover, the RMSEA value exceeded the cut-off criteria for an adequate fit of 0.08.

Item v54 ("Being a housewife is just as fulfilling as working for pay") was identified as the most critical and excluded from the measurement model for the second step of the analysis with the 4-item model. The country-by-country model fit assessment provided an acceptable model fit for 57 countries (the model did not fit the data for Pakistan and Egypt). As with the 5-item model, only partial metric invariance was achieved (Table 3) by releasing two factor loadings; on releasing two intercepts, partial scalar invariance was then tested. However, the results were unsatisfactory, taking into consideration all the global fit measures and the change in model fit from the partial metric model (RMSEA 0.106; ΔRMSEA 0.027; ΔCFI 0.034).

In the third step, because the 4-item model showed a better model fit, this model was tested again while subsequently dropping countries. The gradual selection, carried out on the basis of the modification indices, resulted in dropping 32 countries. Table 4 summarizes the MGCFA results for the remaining 27 countries;[1] partial metric and partial scalar invariance were achieved by releasing two loadings and two intercepts.

1 Azerbaijan; Australia; Bahrain; Armenia; Chile; China; Colombia; Cyprus; Hong Kong; Kazakhstan; South Korea; Kuwait; Lebanon; Libya; New Zealand; Peru; Philippines; Poland; Romania; Russia; Singapore; Slovenia; Spain; Sweden; Trinidad and Tobago; Turkey; United States.

Table 2 MGCFA results. Global fit measures for the exact measurement equivalence of the 5-item model, 57 countries

	Chi2 (dF)	RMSEA	CFI	SRMR
configural	2902.035 (285)***	0.078	0.964	0.032
metric	7763.249 (509)***	0.097	0.900	0.090
partial metric	4007.569 (397)***	0.078	0.950	0.050
partial scalar	6283.398 (453)***	0.093	0.919	0.063

Note: dF= degrees of Freedom; RMSEA= Root Mean Square Error of Approximation; CFI= Comparative Fit Index; SRMR= Standardized Root Mean Square Residual; *** p <0.001; ** p <0.01; * 0.01≤ $p \leq$ 0.1

Table 3 MGCFA results. Global fit measures for the exact measurement equivalence of the 4-item model, 57 countries

	Chi2 (dF)	RMSEA	CFI	SRMR
configural	1469.091 (114)***	0.089	0.979	0.024
metric	3570.189 (282)***	0.088	0.949	0.073
partial metric	1776.035 (172)***	0.079	0.975	0.032
partial scalar	4046.229 (228)***	0.106	0.941	0.056

Note: dF= degrees of Freedom; RMSEA= Root Mean Square Error of Approximation; CFI= Comparative Fit Index; SRMR= Standardized Root Mean Square Residual; *** p <0.001; ** p <0.01; * 0.01≤ $p \leq$ 0.1

Table 4 MGCFA results. Global fit measures for the exact measurement equivalence of the 4-item model, 27 countries

	Chi2 (dF)	RMSEA	CFI	SRMR
configural	575.829 (54)***	0.084	0.982	0.024
metric	1162.631 (132)***	0.075	0.964	0.060
partial metric	1012.997 (105)***	0.079	0.968	0.054
partial scalar	1012.997 (131)***	0.087	0.952	0.060

Note: dF= degrees of Freedom; RMSEA= Root Mean Square Error of Approximation; CFI= Comparative Fit Index; SRMR= Standardized Root Mean Square Residual; *** p <0.001; ** p <0.01; * 0.01≤ $p \leq$ 0.1

Frequentist Alignment Results

The alignment optimization was initially carried out on the original full set of 59 countries. In this first step of the analysis, the overall non-invariance was 50.8% and the Monte Carlo investigation (results for four groups are displayed in Table A.2 in the Appendix) confirmed the poor recovery of the sample; therefore, the alignment results cannot be used to compare means.

This procedure revealed its diagnostic potential. In addition to identifying the overall amount of non-invariance, we immediately recognize the most (non-)invariant parameters. This was the case for item v54 (69 non-invariant parameters), from this point not considered for further analysis, which proceeded in the second step with the 4-item model. The degree of non-invariance dropped to 39.0% and the Monte Carlo investigation confirmed that means comparison would not be reliable, as most of the parameter estimates were biased (Table A.2 in the Appendix).

At this point, the alignment results were used as a diagnostic tool to identify the groups presenting the highest number of non-invariant parameters, which were progressively left out. With a reduced sample of 47 countries, the amount of non-invariance was 26.9%. The results of the Monte Carlo investigation (Table A.3 in the Appendix) displayed a poor replication of the factor means. By excluding countries with more than four non-invariant parameters from the analysis, the use of the alignment procedure with 34 countries[2] provided 21.0% of non-invariance (Table 5). This result met the recommended rule of thumb and could be considered acceptable. The Monte Carlo simulation was run while expecting results as good as those reported by the previous pioneering studies (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2014). While this was not always the case for all the groups and parameters, the global recovery in the Monte Carlo investigation improved, particularly for the factor means that were meant to be compared (Table A.3 in the Appendix). Considering the current state of the art, the results from the alignment optimization are acceptable, even if more simulations designed to determine a clear rule of thumb are probably necessary.

2 Azerbaijan; Bahrain; Armenia; Brazil; Belarus; China; Colombia; Georgia; Ghana; Iraq; Kazakhstan; Jordan; South Korea; Kuwait; Lebanon; Libya; Nigeria; Pakistan; Peru; Philippines; Poland; Qatar; Romania; Russia; Zimbawe; Sweden; Trinidad and Tobago; Tunisia; Turkey; Ukraine; Egypt; Uruguay; Uzbekistan; Yemen.

Table 5 Alignment results. Approximate measurement (non) invariance for intercepts and loadings of the 4-item model, 34 countries

Variable	Intercept	Loadings
V50	31 48 51 (76) (112) 156 170 (268) (288) 368 (398) (400) 410 414 (422) (434) (566) 586 604 608 (616) (634) (642) (643) (716) 752 780 (788) (792) (804) 818 858 (860) (887)	(31) 48 51 76 (112) 156 170 268 288 (368) 398 400 410 (414) 422 434 566 586 604 608 616 634 642 643 716 752 780 788 792 804 (818) 858 860 (887)
V51	31 48 (51) (76) 112 156 (170) 268 288 368 398 400 (410) 414 422 434 566 586 (604) 608 616 (634) (642) 643 716 (752) 780 (788) 792 (804) 818 (858) 860 887	31 (48) 51 76 112 156 170 268 288 368 398 400 410 414 422 434 566 586 604 608 616 (634) 642 643 716 752 780 788 792 804 818 858 (860) 887
V52	31 48 51 76 (112) (156) 170 (268) 288 368 398 400 410 414 422 (434) 566 586 (604) 608 (616) 634 642 643 716 752 780 (788) 792 804 818 858 860 887	31 48 51 76 112 156 (170) 268 288 368 398 400 (410) 414 422 434 (566) 586 (604) (608) 616 634 642 643 716 752 (780) 788 792 804 818 (858) (860) 887
V53	31 48 51 76 112 156 170 268 288 368 398 (400) 410 414 422 434 566 586 604 608 616 (634) 642 643 716 752 780 (788) 792 804 818 858 860 887	31 48 51 76 112 156 170 268 288 368 398 400 410 414 422 434 566 586 604 608 616 634 642 643 716 752 780 788 792 804 818 858 860 887

Note: numbers indicate the country code (see Table 1). The parentheses indicate whether the parameter (intercept or factor loading) is non invariant for that specific group (country code) by variable (v50 to v53).

Table 6 presents the factor means as estimated by the alignment method. The output shows the factor means ordered from the highest (in this case 1.110, for Sweden) to the lowest (-1.242, for Bahrain). The reference codes for each country are given in the second column (and listed in Table 1). Groups with factor means that were significantly different at the 5% level are shown in the last column.

Table 6 Alignment results. 4-item model, factor mean comparison for 34 countries at the 5% significance level in descending order

Ranking	Group	Mean	Groups With Significantly Smaller Factor Mean
1	752 (Sweden)	1.110	604 780 858 170 76 642 616 410 31 716156 804 422 643 398 112 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
2	604 (Peru)	0.590	170 76 642 616 410 716 156 804 422 643 398 112 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
3	780 (Trinidad & Tobago)	0.577	170 76 642 616 410 716 156 804 422 643 398 112 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
4	858 (Uruguay)	0.571	170 76 642 616 410 716 156 804 422 643 398 112 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
5	170 (Colombia)	0.455	76 642 616 410 716 156 804 422 643 398 112 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
6	76 (Brazil)	0.304	642 410 716 156 804 422 643 398 112 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
7	642 (Romania)	0.206	410 716 156 804 422 643 398 112 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
8	616 (Poland)	0.194	410 716 156 804 422 643 398 112 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
9	410 (South Korea)	0.059	716 156 804 422 643 398 112 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
10	31 (Azerbaijan)	0.000	566 414 434 400 860 586 887 818
11	716 (Zimbabwe)	-0.118	643 398 112 268 51 792 288 634 788 368 566 414 434 400 860 586 887 818 48
12	156 (Taiwan)	-0.119	643 398 112 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
13	804 (Ukraine)	-0.135	643 398 112 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
14	422 (Lebanon)	-0.194	643 398 268 51 608 792 288 634 788 368 566 414 434 400 860 586 887 818 48
15	643 (Russia)	-0.307	792 288 634 788 368 566 414 434 400 860 586 887 818 48

Ranking	Group	Mean	Groups With Significantly Smaller Factor Mean
16	398 (Kazakhstan)	-0.318	792 288 634 788 368 566 414 434 400 860 586 887 818 48
17	112 (Belarus)	-0.335	792 288 634 788 368 566 414 434 400 860 586 887 818 48
18	268 (Georgia)	-0.345	792 288 634 788 368 566 414 434 400 860 586 887 818 48
19	51 (Armenia)	-0.369	792 288 634 788 368 566 414 434 400 860 586 887 818 48
20	608 (Philippines)	-0.374	788 368 566 414 434 400 860 586 887 818 48
21	792 (Turkey)	-0.556	788 368 566 414 434 400 860 586 887 818 48
22	288 (Ghana)	-0.573	368 566 414 434 400 860 586 887 818
23	634 (Qatar)	-0.655	566 414 434 400 860 586 887 818
24	788 (Tunisia)	-0.701	566 414 434 400 860 586 887 818
25	368 (Iraq)	-0.801	434 400 860 586 887 818
26	566 (Nigeria)	-0.864	434 400 860 586 887 818
27	414 (Kuwait)	-0.906	887 818
28	434 (Libya)	-1.031	818
29	400 (Jordan)	-1.031	818
30	860 (Uzbekistan)	-1.036	
31	586 (Pakistan)	-1.144	
32	887 (Yemen)	-1.152	
33	818 (Egypt)	-1.184	
34	48 (Bahrain)	-1.242	

Note: In the last column, groups are indicated by the country code (see Table 1)

Sweden, Peru, Trinidad and Tobago, Uruguay, and Colombia proved most supportive of egalitarian gender role attitudes, while Bahrain, Egypt, Yemen, Pakistan, and Uzbekistan ranked lowest of the countries studied. Among the groups dropped, together with the United States, New Zealand, Australia, Palestine, South Africa, Rwanda, India, Algeria, Morocco, Chile, and Ecuador, it is remarkable that most of the European (Cyprus, Estonia, Germany, Netherlands, Slovenia, and Spain), South East Asian (Malaysia, Singapore, and Thailand), and Far Eastern (Japan, Hong Kong, and Taiwan) countries included in this wave of WVS appeared to have a different understanding of the measurement items. These results raise questions for further research: is this because of the culturally different understanding of the

questions and conceptualizing of gender roles? Would adopting a "cluster of cultures approach" (van Vlimmeren et al., 2016) provide further insights?

Concluding Remarks

The current study aimed to contribute to the debate concerning measurement invariance by using data from a large-scale cross-national survey to make applicative use of the frequentist alignment method. Data related to gender role attitudes, and the assessment was addressed to identify the most invariant model across the largest subset of groups (ideally, all). Adopting a step-by-step procedure, both the methods initially led to a model modification by reducing the measurement from a 5-item model to a 4-item model. The two procedures converged in detecting the item v54 ("Being a housewife is just as fulfilling as working for pay") as the least invariant. The option of omitting it found additional support in the critical content analysis of Braun (1998), who pointed out that the understanding of this item can be fairly controversial because of the focus on fulfillment and the benefits from two conditions, rather than on gender roles (Braun, 1998, p. 116).

In the final step, an invariant measurement model was identified for a subset of groups. With the MGCFA, partial scalar invariance was achieved for 27 countries, which would allow for a comparison of means among these countries. However, several model modifications were necessary to achieve it.

On the contrary, with the alignment optimization such modifications are not part of the procedure; the final model retains the same fit of the configural model, which is the best-fitting model possible. By using the frequentist alignment methods, an acceptable degree of non-invariance was achieved for 34 countries, with the rank of the factor means also provided. The results suggest that further substantive work is necessary to understand why the measurement model appears to be equivalent only in this subset of countries, and whether the bias emerges from a culturally different understanding of the questions or from other sources.

The intermediate steps, such as the Monte Carlo investigations, demonstrated that the alignment is not a magic wand, as when the model poorly fits the data, it is evident. Furthermore, the results confirmed the call for caution from Múthen and Asparouhov (2014), such that when the amount of non-invariance is higher than 25%, Monte Carlo investigations are necessary. Nevertheless, further applicative studies are required to establish whether this limit is sufficiently low, and if future studies will be able to rely on it as a clear cut-off criterion without resorting to Monte Carlo investigations.

This study reveals that the alignment procedure is a valuable method to assess measurement equivalence, keeping the good model fit in the most convenient model and allowing factor means comparison for a large number of groups. A possible

further development for the exploration of the alignment method could be a comparison between its use in the frequentist and in the approximate approaches to assess whether the alignment optimization in the Bayesian framework will yield even more promising results than those presented in the current study. At present, only Asparouhov and Muthén (2014) have carried out such a comparison in their simulation study.

References

Alemán, J., & Woods, D. (2016). Value Orientations From the World Values Survey How Comparable Are They Cross-Nationally? *Comparative Political Studies*, *49*(8), 1039–1067. doi:10.1177/0010414015600458

André, S., Gesthuizen, M., & Scheepers, P. (2013). Support for Traditional Female Roles across 32 Countries: Female Labour Market Participation, Policy Models and Gender Differences. *Comparative Sociology*, *12*(4), 447–476.

Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508.

Braun, M. (1998). Gender roles. In Van Deth, JW (Ed.), *Comparative Politics: The Problem of Equivalence*. London, England: Routledge.

Braun, M. (2009). The role of cultural contexts in item interpretation: the example of gender roles. In M. Haller, R. Jowell, & T. W. Smith (Eds.), *The International Social Survey Programme, 1984-2009 : charting the globe* (pp. 395–408). London/New York: Routledge.

Carsey, T. M., & Harden, J. J. (2013). *Monte Carlo Simulation and Resampling Methods for Social Science*. Los Angeles: SAGE Publications.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504.

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. *Quantitative Psychology and Measurement*, *5*, 982. doi:10.3389/fpsyg.2014.00982

Constantin, A., & Voicu, M. (2014). Attitudes Towards Gender Roles in Cross-Cultural Surveys: Content Validity and Cross-Cultural Measurement Invariance. *Social Indicators Research*, *123*(3), 733–751.

Davidov, E. (2010). Testing for comparability of human values across countries and time with the third round of the European Social Survey. *International Journal of Comparative Sociology*, *51*(3), 171–191.

Davidov, E., Cieciuch, J., Meuleman, B., Schmidt, P., Algesheimer, R., & Hausherr, M. (2015). The Comparability of Measurements of Attitudes toward Immigration in the European Social Survey Exact versus Approximate Measurement Equivalence. *Public Opinion Quarterly*, *79*(S1), 244–266.

Davidov, E., Meuleman, B., Billiet, J., & Schmidt, P. (2008). Values and Support for Immigration: A Cross-Country Comparison. *European Sociological Review*, *24*(5), 583–599.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 50–75.

Heath, A., Martin, J., & Spreckelsen, T. (2009). Cross-national Comparability of Survey Attitude Measures. *International Journal of Public Opinion Research*, *21*(3), 293–315.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Inglehart, R., & Baker, W. E. (2000). Modernization, Cultural Change, and the Persistence of Traditional Values. *American Sociological Review*, *65*(1), 19–51.

Inglehart, R., & Welzel, C. (2005). *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. New York: Cambridge University Press.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling, Fourth Edition*. New York: Guilford Publications.

Lomazzi, V. (2017a). Gender role attitudes in Italy: 1988–2008. A path-dependency story of traditionalism. *European Societies*, 1–26. doi:10.1080/14616696.2017.1318330

Lomazzi, V. (2017b). Testing the Goodness of the EVS Gender Role Attitudes Scale. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, Forthcoming. doi:10.1177/0759106317710859

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2017). What to do When Scalar Invariance Fails: The Extended Alignment Method for Multi-Group Factor Analysis Comparison of Latent Means Across Many Groups. *Psychological Methods*. doi:10.1037/met0000113

Moors, G. (2004). Facts and Artefacts in the Comparison of Attitudes Among Ethnic Minorities. A Multigroup Latent Class Structure Model with Adjustment for Response Style Behavior. *European Sociological Review*, *20*(4), 303–320.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335. doi:10.1037/a0026802

Muthén, B., & Asparouhov, T. (2013). BSEM Measurement Invariance Analysis. Mplus Web Notes: No. 17, January 11. (Vol. 17, p. 313). Retrieved February 2, 2017, from https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: the alignment method. *Frontiers in Psychology*, *5*, 978. doi:10.3389/fpsyg.2014.00978

Sjöberg, O. (2004). The Role of Family Policy Institutions in Explaining Gender-Role Attitudes: A Comparative Multilevel Analysis of Thirteen Industrialized Countries. *Journal of European Social Policy*, *14*(2), 107–123.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research*, *25*(1), 78–107.

Stegmueller, D. (2011). Apples and Oranges? The Problem of Equivalence in Comparative Research. *Political Analysis*, *19*(4), 471–487.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthen, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*, 770. doi:10.3389/fpsyg.2013.00770

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *54*(2), 119–135.

van Deth, J. W. (2014). [Review of the book Freedom rising: Human empowerment and the quest for emancipation, by C. Welzel]. *Zeitschrift Für Vergleichende Politikwissenschaft*, *8*(3–4), 369–371.

van Vlimmeren, E., Moors, G. B. D., & Gelissen, J. P. T. M. (2016). Clusters of cultures: diversity in meaning of family value and gender role items across Europe. *Quality & Quantity*, 1–24. doi:10.1007/s11135-016-0422-2

Welzel, C. (2013). *Freedom rising: Human empowerment and the quest for emancipation*. New York: Cambridge University Press.

Welzel, C., & Inglehart, R. F. (2016). Misconceptions of Measurement Equivalence: Time for a Paradigm Shift. *Comparative Political Studies*, *49*(8), 1068–1094.

World Values Survey Association. (2015). WORLD VALUES SURVEY Wave 6 2010-2014 OFFICIAL AGGREGATE v.20150418. (Version file version: WV6_Data_spss_v_2016_01_01 (Spss SAV)). Retrieved from www.worldvaluessurvey.org

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: exact versus approximate measurement invariance. *Frontiers in Psychology*, *6*, 733. doi:10.3389/fpsyg.2015.00733

Appendix

Table A.1 Exploratory Factor analysis results. Extraction Method: Principal Component Analysis

	Full scale	First item excluded
(v49) One of my main goals in life has been to make my parents proud	0.334	
(v50) When a mother works for pay, the children suffer	0.575	0.573
(v51) On the whole, men make better political leaders than women do	0.795	0.796
(v52) A university education is more important for a boy than for a girl	0.694	0.713
(v53) On the whole, men make better business executives than women do	0.820	0.829
(v54) Being a housewife is just as fulfilling as working for pay	0.433	0.435
Initial Eigenvalue	2.415	2.347
% of Variance explained	40.247	46.937

Table A.2 Monte Carlo Simulation for 5-item model and 4-item model. Check of 59 countries Alignment: True values, Estimates, and Coverage (in parentheses). Results for item v50 for the first four groups, N_g=1500.

		5-items model (50,8% of non-invariance)		4-items model (39,0% of non-invariance)	
Group		True value	Estimates (Coverage)	True value	Estimates (Coverage)
1	Loading	0.49	-0.19 (0.00)	0.45	-0.15 (0.00)
	Intercept	2.38	0.16 (0.15)	2.77	-0.01 (0.89)
	Factor Means	1.15	0.21 (0.67)	0.41	0.26 (0.24)
	Factor Variance	0.44	0.74 (0.00)	0.53	0.67 (0.00)
	Residuals variance	0.38	0.00 (0.93)	0.38	0.00 (0.94)
2	Loading	0.41	-0.16 (0.01)	0.41	-0.13 (0.06)
	Intercept	2.11	0.18 (0.22)	2.54	-0.01 (0.95)
	Factor Means	0.02	-0.70 (0.20)	-1.04	-0.45 (0.32)
	Factor Variance	0.26	0.43 (0.00)	0.25	0.29 (0.33)
	Residuals variance	0.56	0.00 (0.94)	0.57	0.00 (0.93)
3	Loading	0.32	-0.12 (0.00)	0.27	-0.09 (0.01)
	Intercept	2.32	0.10 (0.18)	2.55	-0.01 (0.95)
	Factor Means	0.28	-0.31 (0.36)	-0.52	-0.21 (0.39)
	Factor Variance	0.52	0.88 (0.00)	0.59	0.77 (0.00)
	Residuals variance	0.52	0.88 (0.00)	0.71	0.00 (0.96)
4	Loading	0.25	-0.09 (0.06)	0.22	-0.08 (0.19)
	Intercept	2.07	0.08 (0.51)	2.27	0.00 (0.95)
	Factor Means	0.85	-0.01 (0.92)	0.05	0.07 (0.80)
	Factor Variance	0.30	0.50 (0.00)	0.34	0.45 (0.00)
	Residuals variance	0.65	0.00 (0.97)	0.65	0.00 (0.94)

Table A.3 Monte Carlo Simulation for 4-item model. Check of 47 and 34 countries Alignment: True values, Estimates, and Coverage (in parenthesis). Results for item v50 for the first four groups, N_g=1500.

		4-items model 47 countries (26,9% of non-invariance)		4-items model 34 countries (21,0% of non-invariance)	
Group		True value	Estimates (Coverage)	True value	Estimates (Coverage)
1	Loading	0.30	0.03 (0.96)	0.28	-0.03 (0.77)
	Intercept	2.61	-0.20 (0.43)	2.47	-0.08 (0.73)
	Factor Means	-1.64	0.77 (0.22)	-1.24	0.16 (0.90)
	Factor Variance	0.47	-0.08 (0.76)	0.53	0.15 (0.96)
	Residuals variance	0.57	0.00 (0.93)	0.57	0.11 (0.96)
2	Loading	0.22	0.01 (0.98)	0.16	-0.03 (0.68)
	Intercept	2.58	-0.12 (0.32)	2.86	-0.03 (0.72)
	Factor Means	-0.79	0.54 (0.23)	-0.34	0.19 (0.61)
	Factor Variance	0.92	-0.08 (0.80)	1.08	0.45 (0.41)
	Residuals variance	0.71	0.00 (0.94)	0.66	0.00 (0.92)
3	Loading	0.18	0.01 (0.91)	0.40	-0.06 (0.50)
	Intercept	2.30	-0.10 (0.38)	2.44	-0.09 (0.59)
	Factor Means	-0.10	0.53 (0.27)	-0.80	0.10 (0.84)
	Factor Variance	0.55	-0.05 (0.81)	0.78	0.36 (0.35)
	Residuals variance	0.65	0.00 (0.98)	0.47	0.00 (0.96)
4	Loading	0.16	0.01 (0.95)	0.29	-0.05 (0.49)
	Intercept	2.92	-0.08 (0.34)	1.86	-0.06 (0.56)
	Factor Means	-0.73	0.52 (0.25)	-1.03	0.07 (0.64)
	Factor Variance	1.07	-0.10 (0.73)	0.85	0.35 (0.56)
	Residuals variance	0.66	-0.01 (0.90)	0.48	0.00 (0.95)

Table A.4 Mplus input excerpts for Fixed alignment ML estimation for the 4-item model in 34 countries

```
TITLE:      WVS 6 gender roles alignment;
DATA:       file is WV6_gender role.dat;
VARIABLE:   Names are
              V2 v50 v51 v52 v53 v54;
            usevariables are
              v50 v51 v52 v53;
              missing = all (999);
              classes= c(34);
              knownclass is c(v2=31 v2=48 v2=51 v2=76 v2=112 v2=156 v2=170
            v2=268 v2=288 v2=368 v2=398 v2=400 v2=410 v2=414 v2=422
            v2=434 v2=566 v2=586 v2=604 v2=608 v2=616 v2=634 v2=642
            v2=643 v2=716 v2=752 v2=780 v2=788 v2=792 v2=804 v2=818
            v2=858 v2=860 v2=887);
ANALYSIS:   type = mixture;
             estimator=ML;
             alignment=fixed;
MODEL:      %overall%
              GI by v50 v51 v52 v53;
OUTPUT:     align stand Tech1 Tech8;
```

Table A.5 Mplus input excerpts Monte Carlo for simulation for the 4-item model in 34 countries

```
TITLE:      WVS 6 gender roles alignment MC1;
DATA:       file is WV6_gender role.dat;
VARIABLE:   Names are
              V2 v50 v51 v52 v53 v54;
            usevariables are
            v50 v51 v52 v53;
            missing = all (999);
            classes= c(34);
            knownclass is c(v2=31 v2=48 v2=51 v2=76 v2=112 v2=156 v2=170
            v2=268 v2=288 v2=368 v2=398 v2=400 v2=410 v2=414 v2=422
            v2=434 v2=566 v2=586 v2=604 v2=608 v2=616 v2=634 v2=642
            v2=643 v2=716 v2=752 v2=780 v2=788 v2=792 v2=804 v2=818
            v2=858 v2=860 v2=887);
ANALYSIS:   type = mixture;
            estimator=ML;
            alignment=fixed;
MODEL:      %overall%
            GI by v50 v51 v52 v53;
OUTPUT:     Tech1 svalues;
```

```
TITLE:      WVS 6 gender roles alignment MC simulation;
montecarlo: names = v50 v51 v52 v53 v54;
            ngroups=34;
            nobservations=34(1500);
            nreps= 100;
            repsave=all;
            save=n1500f-22rep*.dat;
analysis:   type=mixture;
            estimator=ML;
            alignment=fixed (22);
            processors=8;
```

```
model population:
%overall%
  gi by v50 -v53*1;
%G#1%
  gi BY v50*0.44755;
  gi BY v51*0.66271;
  gi BY v52*0.41177;
  gi BY v53*0.68205;
  [ v50*2.39376 ];
  [ v51*2.10195 ];
  [ v52*2.75848 ];
  [ v53*2.01374 ];
  [ gi*0 ];
  v50*0.57993;
  v51*0.36809;
  v52*0.71822;
  v53*0.32406;
  gi*1;
%G#2%
  […]

Model:
%overall%
  gi by v50 -v53*1;
%G#1%
  gi BY v50*0.44755;
  gi BY v51*0.66271;
  gi BY v52*0.41177;
  gi BY v53*0.68205;
  [ v50*2.39376 ];
  [ v51*2.10195 ];
  [ v52*2.75848 ];
  [ v53*2.01374 ];
  [ gi*0 ];
  v50*0.57993;
  v51*0.36809;
  v52*0.71822;
  v53*0.32406;
  gi*1;

 %G#2%
  […]
```

methods, data, analyses | Vol. 12(1), 2018, pp. 105-126 DOI: 10.12758/mda.2017.08

Effects of Rating Scale Direction Under the Condition of Different Reading Direction

Dagmar Krebs[1] *& Yaacov G. Bachner*[2]

[1] *Justus Liebig University, Giessen, Germany,*
[2] *Ben-Gurion University of the Negev, Beer-Sheva, Israel*

Abstract

Because response scales serve as orientation for respondents when mapping their answers to response categories, it can be expected that the decremental (from positive to negative) or incremental (from negative to positive) order of a response scale provides information that influences response behavior. If respondents interpret the first category on a scale as signifying "most accepted," then starting an agree/disagree scale with "agree completely" or "disagree completely" may result in their forming different subjective hypotheses about the "most acceptable" response. If this principle applies in general, respondents' reactions to horizontal response scales with different orders of response categories should be similar in the two directions of reading – right to left or left to right. This paper tests two hypotheses: first, that decremental scales elicit more positive responses than incremental scales; second, that this pattern holds under the condition of different reading direction. These hypotheses were tested using a German and an Israeli student sample. Seven-point decremental and incremental scales were applied in each sample; only the scale endpoints were verbally labeled. The questions asked related to extrinsic and intrinsic job motivation and achievement motivation. For data collection, a split-ballot design with random assignment of respondents to decremental and incremental scales was applied in both samples. Results revealed that response-order effects occur similarly in the right-to-left and the left-to-right reading direction.

Keywords: response-order effect; scale direction; reading direction; primacy and recency effect; satisficing

Introduction

In this article, we investigate whether response-order effects occur similarly in different reading directions (i.e., right to left vs. left to right). For this comparison, we conducted an experiment in Israel and Germany using rating scales. As response-order effects, we investigated the effects of scale direction on response behavior by applying a decremental (i.e., from positive to negative) and an incremental (i.e., from negative to positive) response scale.

Since the beginning of attitude measurement, social scientists have defined attitudes as evaluations expressing the degree of favorableness toward an attitude object. Therefore, attitude measurement relies on responses expressing this degree of favorableness on a continuum extending from favor to disfavor, agree to disagree, etc. The use of rating scales in social science surveys has a decades-long tradition. Information retrieved from scale handbooks (Bruner, 2013; Fowler, 1995; Robinson, Shaver and Wrightsman, 1999) shows that over 90 per cent of attitude measurement used the rating scale technique developed by Likert (1932). This technique originally applied a five-point, fully labeled scale offering response categories on an approve/disapprove continuum with a neutral midpoint (i.e., strongly approve, approve, undecided, disapprove, strongly disapprove). Since these early days, a vast amount of methodological research has investigated the effects of different response-scale attributes on response behavior.

With respect to the effect(s) of scale length, Miller (1956), in an intriguing article titled "The Magical Number Seven Plus or Minus Two," reviewed research suggesting that respondents have the capacity to process seven response categories (plus or minus two). Since then, there have been numerous recommendations for the optimal number of scale points (e.g., Alwin, 1997, 2007; Krosnick & Fabrigar, 1997; Kieruj & Moors, 2010; Krosnick & Presser, 2010; Preston & Colman, 2000; Saris & Gallhofer, 2014; Weng, 2004). Although results of empirical studies are somewhat inconclusive with regard to the optimal number of response categories, there seems to be some consensus that more response categories yield more information about the variable of interest (Revilla, Saris, & Krosnick, 2014). Whereas rating scales with too few categories may fail to discriminate between respondents with different underlying judgments, too many categories may make it impossible for respondents to distinguish reliably between adjacent categories. An extensive overview of the literature on (unipolar and bipolar) scales revealed that bipolar scales with around seven points, and unipolar scales with between five and seven points, yielded greatest measurement reliability (Alwin, 2007; Krosnick & Fabrigar, 1997) and therefore seem to represent the best compromise. An over-

Direct correspondence to
Dagmar Krebs, Justus Liebig University, Giessen, Germany
E-mail: dagmar.krebs@sowi.uni-giessen.de

view of 603 scales used in questionnaires revealed that 55% used a 7-point scale and 30% used a 5-point scale (Weijters, Cabooter, & Schilleweart, 2010). When it comes to the complexity of scales with seven compared to five response categories, there is consensus within the scientific (survey methodology) community that high-educated respondents can handle more differentiated scales, but that 5-point scales should be used in general population surveys (Weijters, et al. 2010). Therefore, as we were using student samples, we decided to employ a 7-point scale in our experimental study.

Empirical research on completely or partially labeled response scales has been published by Krosnick (1999); Krosnick and Fabrigar (1997); Krosnick and Presser, 2010; Tourangeau, Rips, and Raisinski (2000); and Weng (2004). Krosnick and Fabrigar (1997) and Menold and Bogner (2014) expressed a preference for completely labeled scales, arguing that verbal labels offered greater clarity of response alternatives, especially for respondents with a low level of education. However, the authors admitted that a 7-point scale with refined verbal labels for each response category could be more demanding than a 7-point scale with verbal labels only at the endpoints. Formulating (seven) verbal labels is difficult enough in one language. However, it is even more challenging when, as in our study, two languages are used (i.e., Hebrew and German). According to Fowler and Cosenza (2008), numbers between the verbally labeled endpoints translate much better across languages than do adjectives. We therefore decided to employ 7-point scales with verbal labels at the endpoints and numbers in between.

Regarding scale polarity, we refer to the findings of Schwarz, Knäuper, Hippler, Noelle-Neumann, and Clark (1991), who compared two sets of a 10-point rating scale with bipolar verbal endpoints. One set contained numerical values from -5 to +5, whereas values of the other set ranged from 0 to 10. Regardless of the scale labels, responses piled up in the positive half of the scale. Apparently, the negative numbers changed the meaning of the verbal labels, thereby suggesting that respondents interpreted the endpoints not as logical complements but as polar opposites (success/no success vs. success/failure). To avoid this unintended effect, there is a tendency in the literature to use unipolar scales (Schaeffer & Presser, 2003). Accordingly, the endpoint-labeled, 7-point scale used in this study is unipolar.

Response-order effects also have quite a long tradition in research on survey methodology. Empirical results relating to response-order effects for categorical response options (Sudman, Bradburn, & Schwarz, 1996) extend to rating scales with ordered categories (Bishop & Smith, 2001; Krosnick & Alwin, 1987; Krosnick, Narayan, & Smith, 1996; Malhotra, 2008; Schwarz, Hippler, & Noelle-Neumann, 1992; Yan & Keusch, 2015). From these studies, it seems obvious that response-order effects occur both in categorical scales and in rating scales, but that these effects are much less pronounced in rating scales (Sudman, Bradburn, & Schwarz, 1996). The absence of consensus on the order or direction of rating scales

might be due to the (comparatively) small response-order effects in rating scales. The decision whether a response scale should start with the positive or the negative response category seems to be largely up to the individual researcher. This circumstance applies within the left-to-right reading direction. However, little to nothing is known about response-order effects in another reading direction, namely right to left. Although Rayner (1998) mentioned the possibility that writing/reading direction influences response behavior, there has been no systematic research on the occurrence of response-order effects in the right-to-left versus the left-to-right reading/writing direction. This lack of research prompted us to conduct an experiment on response-order effects with Israeli (reading right to left) and German (reading left to right) respondents using a decremental scale (from positive to negative) and an incremental scale (from negative to positive). Both scales were 7-point, endpoint labeled, and unipolar.

Theoretical Background and State of Research

The existence of response-order effects has been known since the beginning of survey methodology in the 1920s (Mathews, 1929) and has been shown in many empirical studies (e.g., Bishop & Smith, 2001; Malhotra, 2008; Yan & Keusch, 2015). As Krosnick and Alwin (1987) demonstrated, the shape of these response-order effects depends to a large extent on presentation mode – auditory or visual. Whereas the auditory mode promotes recency effects (i.e., endorsement of response alternatives appearing late on a list or a response scale), visual presentation promotes primacy effects (i.e., endorsement of alternatives appearing early on a list or a response scale). Although it was well-known for years that the order in which response alternatives are presented to respondents can significantly alter the results and conclusions of public opinion polls (Bishop & Smith, 2001, p. 479), a theoretical explanation for this phenomenon was lacking. As recently as the 1980s, two complementary explanations were offered: satisficing theory, proposed by Krosnick and Alwin (1987) and Krosnick (1991), and cognitive elaboration theory, proposed by Sudman, Bradburn, and Schwarz (1996). Both theories explain the occurrence of primacy effects by deeper cognitive processing of response alternatives presented earlier rather than later in a list. And both theories also emphasize mode differences and expect primacy effects in the case of visual presentation and recency effects in the case of auditory presentation. Furthermore, empirical evidence generally shows smaller effects for rating scales than for categorical scales (Sudman et al., 1996).

Despite these similarities, there are some differences between the two theories in terms of their perspective on cognitive processes that result in response-order effects. Satisficing theory is based on the principle of rational choice, whereby decision makers in possession of limited information try to find an adequate rather than

an optimal solution. This approach, known as "bounded rationality" or "satisficing" (Simon, 1959), explains response-order effects as a strategy to minimize cognitive effort, which, following Krosnick (1991), results in a primacy effect in visual presentations. Accordingly, the primacy effect occurs either because respondents select the first acceptable response category, thereby inhibiting consideration of later ones, or because they are not capable of processing all the response categories equally, thereby leading to preferential selection of the initial ones. With respect to response-order effects, Krosnick (1992) and Krosnick, Narayan, and Smith, (1996) describe this response behavior as "weak satisficing" that leads respondents to select the first acceptable response alternative on a response scale.

Cognitive elaboration theory, by contrast, is based on cognitive processes similar to those that occur in persuasive communication. Hence, one can conceive of the "measurement unit" comprising a question and a response scale as a short persuasive argument that elicits positive or negative cognitive responses (Sudman, Bradburn, & Schwarz, 1996). From this perspective, if recipients develop positive associations with the "message," then positive attitude change will occur. However, if recipients develop negative associations with the "message," they will back away from it. Transferred to response scales, this principle implies that a "measurement unit" that offers positive response alternatives first (e.g., in a decremental scale) draws respondents toward a positive response, whereas a "measurement unit" that offers negative response alternatives first (e.g., in an incremental scale) draws respondents away from the negative response. Based on this consideration and the postulate that, in visual presentation, it is easier to cognitively elaborate response categories at the beginning of a list than categories appearing later, cognitive elaboration theory can predict a primacy effect for decremental scales and a recency effect for incremental scales.

To sum up: Whereas satisficing theory can explain the occurrence of a primacy effect, cognitive elaboration theory can explain, in addition, a recency effect for incremental response scales. Therefore, in combination, these two theories enable us to formulate differentiated expectations for effects associated with decremental and incremental response scales.

Studies investigating the effects of response order (Krebs & Hoffmeyer-Zlotnik, 2010; Krebs, 2012; Krosnick & Alwin, 1987; Krosnick, 1991) have shown that response-order effects are observable for different content areas and different samples. However, effects of response order are chronically small, especially when the response scale is presented horizontally (Höhne & Lenzner, 2015; Menold & Bogner, 2014).

The results of these studies support both satisficing theory and cognitive elaboration theory. However, they were conducted in cultures in which the left-to-right reading direction prevails. If the theoretical considerations describe a general principle of the response process, then the results of response-order effects should

be replicable in a different cultural context with a right-to-left reading direction. Therefore, we investigate the effect of scale direction within the right-to-left and the left-to-right reading direction by comparing responses on decremental and incremental scales in an Israeli and a German group of respondents.

Hypotheses

As a global hypothesis, we postulate that response-order effects occur due to scale direction, and that these effects are similarly observable in the right-to-left and the left-to-right reading directions.

We derive our hypotheses on response-order effects from satisficing theory and cognitive elaboration theory. Both theories predict primacy effects for response alternatives that appear first on a scale. However, cognitive elaboration theory predicts that this primacy effect will occur primarily on decremental scales. Therefore, we expect higher proportions of responses at the beginning of decremental response scales than at the beginning of incremental response scales (hypothesis 1).

According to satisficing theory, one would expect a primacy effect also in the case of an incremental response scale. However, based on persuasive-communication reasoning, cognitive elaboration theory predicts that a recency effect for incremental scales is more likely than a primacy effect, because the negative response alternatives, although presented early on the response scale, elicit negative cognitive associations and are therefore less likely to receive endorsement. Taking into account (a) "positivity bias" (Tourangeau, Rips, & Raisinski, 2000), which describes respondents' preference for positive answers, and (b) satisficing theory, which implies that respondents engage in "weak satisficing" by looking for the first acceptable response alternative on a response scale, piling of responses on incremental scales is likely to occur on the middle to positive response alternatives. Therefore, compared to decremental response scales, in the case of incremental scales we expect higher proportions of responses near the middle of the scale (hypothesis 2).

We expect, further, that this "retreat to the middle of the scale" in the case of the incremental scales will be observable in the means. In line with Toepoel, Das, and van Soest (2009), we expect that positive responses will occur more often on decremental than on incremental scales. Because the number of less positive answers is higher on incremental scales (Krebs & Hoffmeyer-Zlotnik, 2010), and all values were coded from 1 (positive) to 7 (negative), we expect to observe lower means (more positive answers) on decremental than on incremental scales (hypothesis 3).

Methods

Survey Questions

To study response-order effects in different reading directions, we adapted 12 items from the Cross Cultural Survey for Work and Gender Attitudes 1991-2010 (Frieze, 2010) and the German General Social Survey (ALLBUS) 2006. This approach has the advantage of using repeatedly tested questions. Four of these questions refer to extrinsic job motivation, four to intrinsic job motivation, and four to achievement motivation.

Extrinsic job motivation refers to the importance of anticipated job characteristics (e.g., income and career prospects) that are not primarily under an individual's control. Intrinsic job motivation refers to job commitment (e.g., autonomy and responsibility). Achievement motivation refers to "competitiveness," and implies a preference for interpersonal challenges. Whereas intrinsic job motivation and achievement motivation describe attitudes toward a job or toward possible competitors, extrinsic job motivation describes requirements that job characteristics should meet (Krebs, Berger, & Ferligoj, 2000; Spence & Helmreich, 1983).

The decision to use motivational questions for this study is based on the authors' experience of (nearly) identical results: Achievement motivation, intrinsic job motivation, and extrinsic job motivation proved to be stable across different student cohorts and over time. The motivational questions in the questionnaire were followed by several questions on political and societal issues (which are not the subject of this paper).

All questions were presented in grids[1] with a unipolar 7-point response scale with numeric values between the verbal endpoints (see Appendix for the questions used). Achievement motivation was measured on a scale similar to an agree/disagree response scale, whereas (extrinsic and intrinsic) job motivation was assessed on an importance scale. Accordingly, the verbal endpoints on the decremental scales were (a) *applies to me completely* (=1) and *does not apply to me at all* (=7)

1 Although grids have been criticized for their disadvantages relating to the "manner of question asking" (Höhne & Krebs, 2017), this question format is still very popular and widely used in surveys. Grids allow parsimonious presentation, which relates directly to questionnaire production costs. Moreover, by using this question format in our study, all questions could be printed on the front and back of just one sheet of paper, thereby avoiding discouraging respondents by presenting them with a multi-page questionnaire. Only recently have item-specific (IS) question formats been discussed as an alternative to agree-disagree (A/D) questions in grids. However, empirical evidence that IS questions yield better data quality has yet to be confirmed.

and (b) *very important* (=1) and *not important at all* (= 7).[2] On the incremental scales, the endpoints were labeled inversely. Response scales were presented horizontally next to the items and either on the right side (German) or on the left side (Hebrew), depending on the reading direction. Whereas the Cross Cultural Survey for Work and Gender Attitudes 1991–2010 (Frieze, 2010) used 5-point scales, ALLBUS 2006 used 7-point scales. As we were using student samples, and as high-educated respondents can handle more differentiated scales, we decided to employ 7-point scales.

Data Collection

The study took place in spring and early summer 2008 at the Justus-Liebig-University in Giessen, Germany and Ben-Gurion University of the Negev, in Beer-Sheva, Israel. It was designed as a split-ballot experiment. Respondents were students from the pedagogical or public health department, who were not familiar with social science methodology. Questionnaires were distributed and completed during lectures. To ensure randomization of split versions (decremental vs. incremental scale), questionnaires were sorted systematically before distribution. All students in the lecture hall were invited to participate, and received and completed the questionnaire. Students were informed that they were participating in a study on the quality of survey questions; confidentiality was assured.

Items and item sequence were identical in the two split versions. Only the direction of response scales in the splits varied. The questionnaire took about 10 minutes to complete, and questionnaires were collected immediately after completion.

In all, we obtained 175 completed questionnaires in Israel and 250 in Germany. In Germany, the questionnaire with the decremental scales was completed by 115 respondents (78% female), and the questionnaire with the incremental scales was completed by 105 respondents (75% female); 30 respondents did not answer the gender question. In Israel, the questionnaire with the decremental scales was completed by 62 persons (68% female); 113 persons (74% female) completed the questionnaire with the incremental scales.

2 Although items were adapted from the Cross Cultural Survey for Work and Gender Attitudes 1991–2010 (Frieze, 2010), response scales were not. That study did not measure the importance of job characteristics but rather agreement/disagreement with items such as "It is important to me that …."

Data Analyses

Because we were interested in response-order effects associated with decremental and incremental scales in the left-to-right (German group) and the right-to-left (Israeli group) reading directions, we conducted the analyses for each group separately. Comparing results from response scales with different directions is not possible without assessing the measurement equivalence of the two scales. To ensure measurement equivalence, we followed four steps of hierarchical modeling (Revilla, 2013): First, a confirmatory factor analysis (CFA) model was specified separately for the decremental and the incremental scales in the German and the Israeli group. Second, the CFA was conducted simultaneously to test for configural invariance of the decremental and the incremental scales within each group. In the third and fourth steps, respectively, metric invariance was tested by restricting the factor loadings to equality, and scalar invariance was tested by additionally restricting the intercepts to equality. A meaningful comparison of latent means is possible only if scalar invariance holds. Because all indicators were measured on a 7-point scale, we assumed continuous scale level. For all analyses, we used Mplus version 6.12 (Muthén & Muthén, 1998-2010) and applied the MLM discrepancy function, thereby allowing for non-normality of distributions (Byrne, 2012).

However, before comparing the latent means of achievement motivation, intrinsic job motivation, and extrinsic job motivation, we inspected the empirical distribution parameters and the proportions of positive and negative responses on the decremental and incremental scales.

Results

To ensure unequivocal statistical analyses, all values were coded from 1 (positive) to 7 (negative). First, we inspected the parameters of the empirical distributions for all indicators. Then we tested for measurement equivalence between decremental and incremental scales within the right-to-left and the left-to-right reading directions. Next, we compared proportions of the empirical referents (unweighted sum scores) of achievement motivation, intrinsic and extrinsic job motivation within each reading direction. And finally, we compared the latent means of the three motivational constructs.

Descriptive Statistics

From Table 1, it is obvious that the decremental scales, in particular, have extreme kurtosis values (bolded) in both reading directions. This is observable especially in the case of two intrinsic job motivation indicators (applying skills and realizing

Table 1 Means, standard deviations, skewness, and kurtosis for decremental and incremental scale directions within the right-to-left (Hebrew) and the left-to-right (German) reading directions

	Decremental Order				Incremental Order			
Survey Questions	Mean	SD	Skew-ness	Kurto-sis	Mean	SD	Skew-ness	Kurto-sis
Reading: Right to Left (Hebrew)								
Enjoy competition	4.16	1.79	-0.10	-1.13	4.57	1.62	-0.29	-0.80
Important to be better	2.86	1.68	0.92	-0.22	3.47	1.77	0.39	-0.98
Enjoy being better	3.23	1.80	0.70	-0.57	3.74	1.72	0.33	-1.04
Spurred on by competition	3.19	1.73	0.68	-0.39	3.43	1.61	0.37	-0.76
Autonomy	2.10	1.16	1.57	3.69	2.27	1.10	0.99	1.87
Applying skills	**1.61**	**1.09**	**2.73**	***8.94***	1.59	0.76	0.94	-0.26
Responsibility	2.15	1.24	1.41	2.41	2.20	1.14	0.80	-0.22
Realizing ideas	**1.76**	**1.07**	**2.32**	***7.64***	1.75	1.01	1.50	2.39
Income	2.21	1.19	1.33	1.82	2.25	1.21	1.10	1.01
Prospects	**1.84**	**0.85**	**1.86**	***6.91***	1.97	0.92	0.73	0.02
Career	1.84	0.83	0.63	-0.46	2.02	0.93	0.70	-0.02
Respect	1.81	0.85	1.33	2.37	1.93	1.02	0.83	-0.30
Reading: Left to Right (German)								
Enjoy competition	3.84	1.65	0.46	-0.70	3.81	1.69	0.18	-1.04
Important to be better	3.71	1.65	0.45	-0.87	4.03	1.65	0.09	-0.96
Enjoy being better	4.00	1.73	0.12	-1.08	4.28	1.86	-0.04	-1.24
Spurred on by competition	3.25	1.73	0.58	-0.67	3.27	1.64	0.72	-0.18
Autonomy	1.99	0.93	0.83	0.18	2.08	1.15	1.32	2.33
Applying skills	**1.60**	**0.86**	**2.46**	***10.69***	**1.68**	**0.93**	**2.17**	***7.91***
Responsibility	1.99	1.03	1.16	1.30	2.05	1.02	0.75	-0.17
Realizing ideas	**1.72**	**0.92**	**2.01**	***7.28***	1.80	0.92	1.29	1.55
Income	2.80	1.10	0.84	1.20	2.96	1.33	0.90	1.01
Prospects	3.31	1.37	0.49	-0.21	3.45	1.45	0.37	-0.43
Career	3.39	1.44	0.51	-0.32	3.52	1.58	0.27	-0.77
Respect	3.13	1.41	1.05	0.73	3.08	1.46	0.76	0.12

Notes: The first four questions refer to achievement motivation, the next four questions refer to intrinsic job motivation, and the last four questions refer to extrinsic job motivation.

ideas). These extremes occurred in both groups, their exclusion would have considerably minimized the number of indicators for intrinsic job motivation, and the MLM discrepancy function took non-normality into account. Therefore, to ensure comparability, all variables were retained in the analyses for both groups.

Inspecting the item means in Table 1 more closely reveals higher values on the incremental scales than on the decremental scales. According to the coding from 1 (positive) to 7 (negative), this indicates that responses in both reading directions are slightly but systematically more negative on the incremental scales. This similarity of distributions corresponds to our global hypothesis that direction effects are the same in the right-to-left (Hebrew) and left-to-right (German) reading directions.

Although comparing proportions (as described in hypotheses 1 and 2) would belong in the present section, we prefer first to ensure measurement equivalence of the decremental and incremental scales within each reading direction, and to postpone comparisons between scale directions.

Measurement Equivalence

First, we formulated a first-order confirmatory factor analysis (CFA) model with three latent variables (achievement motivation, intrinsic job motivation and extrinsic job motivation) and four indicators each. This (baseline) model was tested within the German and the Hebrew reading directions for the decremental and incremental scales separately. For the German group, values of modification indices (MI) together with expected parameter change (EPC) values suggested the inclusion of two residual covariances, one for the two items of achievement motivation referring to “competition” and one for the two items of extrinsic job motivation referring to “career prospects” and “promotion prospects” (see Appendix for item wording). Because of their obvious overlap in item content, these two residual covariances were included in the model for the German group. Furthermore, for the Israeli group, one cross-factor loading from intrinsic job motivation to “income” (extrinsic job motivation item) suggested by MI and EPC values was included in the model. To ensure comparability between scale directions within each group, these model re-specifications were applied to the decremental and the incremental baseline models. In the German group, the model for the decremental scales had a comparative fit index (CFI) of 0.978 and a root mean square error of approximation (RMSEA) of 0.050, whereas the incremental model had a CFI of 0.972 and an RMSEA of 0.048. In the Israeli group, the global fit measures had CFIs of 1.00 and 0.992 and RMSEAs of 0.00 and 0.027, respectively, for the decremental and incremental scales. Because we are investigating response-order effects due to scale direction within the left-to-right (German) and the right-to-left (Israeli) reading directions, we continued by testing configural, metric, and scalar invariance of the re-specified models for the decremental and incremental scales within each

Table 2 Testing measurement equivalence of decremental and incremental response order in the right-to-left (Hebrew) and the left-to-right (German) reading directions for the model containing three latent variables with four indicators each

	χ^2	*df*	χ^2-Diff.	CFI	RMSEA
*Right to Left (Hebrew)**					
Configural	103.39 (1.13)	100		0.996	0.020
Metric	121.19 (1.15)	110	16.69	0.986	0.034
Scalar	131.05 (1.14)	122	4.61	0.988	0.029
Means	125.78 (1.14)	119		0.991	0.026
*Left to Right (German)***					
Configural	127.64 (1.15)	98		0.975	0.049
Metric	144.00 (1.19)	107	15.12	0.969	0.053
Scalar	150.80 (1.17)	119	5.12	0.974	0.046
Means	148.31 (1.17)	116		0.973	0.047

Notes: * Model with one cross loading. ** Model with two residual covariances. Values in brackets are scale correction values required for MLM based model comparisons by way of chi-square difference testing, see χ^2-Diff. for the respective values of χ^2-differences between models.

group. Table 2 shows the results. According to Byrne (2012) and Revilla (2013), the decision of invariance can be based on the difference in CFI and RMSEA values between the configural, metric, and scalar invariance models. With respect to this criterion, a change in CFIs greater than 0.01 accompanied by a change in RMSEA greater than 0.015 would be indicative of non-equivalence. As can be seen from Table 2, differences in the CFI and RMSEA values between the configural, metric, and scalar invariance models meet this criterion: The largest difference for the CFIs is 0.01 and for the RMSEAs is 0.014 Additionally, chi-square differences are not significant. Therefore, because scalar invariance for the decremental and incremental scales is supported by the data within the German and the Israeli groups, measurement invariance can be accepted, and comparison of latent means based on scale directions is possible. The last row of Table 2 refers to the global fit statistics for the model comparing latent means described in section Comparison of Latent Means.

First, however, we compare proportions of negative and positive answers in the two scale directions.

Comparison of Proportions

For the comparison of proportions, we calculated unweighted sum scores for achievement motivation, intrinsic job motivation and extrinsic job motivation. These sum scores (divided by the number of items constituting the sum) are grouped into three blocks referring to positive (response categories 1 and 2), middle (response categories 3, 4, and 5), and negative answers (response categories 6 and 7). For easier reading, the ratios of response proportions in these three blocks on decremental versus incremental scales are computed; they give an impression of how respondents reacted to different scale directions. Distributions of responses on decremental and incremental scales are very similar within the left-to-right (German) and the right-to-left (Hebrew) reading directions as revealed by the results of a chi-square test. Hypotheses 1 and 2 were tested by means of Fisher's exact test. Results are presented in Table 3, and all values are coded in the 1 (positive) and 7 (negative) direction.

All distributions in Table 3 show the typical, well-known pattern that respondents dislike both the extreme positive and the extreme negative categories on a response scale. In the case of achievement motivation, for example, response proportions on the decremental scale are 23% higher on the second response category than on the first response category in the Israeli group, and they are 13% higher in the German group. Likewise, response proportions in the Israeli group are 7% higher on the second-last category of the incremental scale (the second response category in the questionnaire) than on the extreme category; in the German group, they are 5% higher. As postulated in hypothesis 1, proportions of positive answers at the beginning (response categories 1 and 2) of the decremental response scales are higher than those on the incremental response scales. This hints to a primacy effect, which is observable in the ratio of positive responses between scale directions (29% vs. 15% in the Israeli group and 17% vs. 15% in the German group). Responses on the incremental scales start piling in the middle (response categories 3, 4 and 5) with a ratio of decremental to incremental of 63% vs. 69% for the Israeli group and 69% vs. 75% for the German group). Furthermore, for all sum scores, proportions on the middle response categories of the incremental scales are higher than on middle response categories of the decremental scales, thereby confirming that respondents tend to back away from the negative response. For intrinsic job motivation and in the case of the incremental scales it can be observed in both (German and Israeli) groups that proportions of responses increase from the negative toward the positive end of the scale, thereby leaning toward a recency effect (hypothesis 2). The same pattern occurs for extrinsic job motivation in the Israeli group. Altogether, the postulated differences between the decremental and the incremental scale directions, although observable both in the Hebrew and the German reading directions, are not significant. Because the observed patterns of

Table 3 Proportions of positive, middle, and negative responses for achievement motivation, intrinsic job motivation and extrinsic job motivation (unweighted sum scores) on decremental and incremental scales within the right-to-left (Hebrew) and the left-to-right (German) reading directions

	Reading Right to Left (Hebrew)			*Reading Left to Right (German)*		
	decremental	incremental	ratio decr.: incr.	decremental	incremental	ratio decr.: incr.
Achievement	%	%	%	%	%	%
Applies to me completely	3	5	29:15	2	3	17:15
2	26	10		15	12	
3	21	26	63:69	33	19	69:75
4	31	27		19	32	
5	11	16		17	24	
6	5	12	8:17	9	8	15:11
Does not apply to me at all	3	5		6	3	
	$\chi^2(6)=10.62$, n.s.			$\chi^2(6)=13.38$, p=0.04		
Intrinsic						
Very important	31	29	81:76	34	30	85:78
2	50	47		49	48	
3	13	19	18:25	15	16	18:21
4	5	4		2	5	
5	0	2		1	0	
6	0	0	2:0	0	0	1:1
Not important at all	2	0		1	1	
	$\chi^2(5)=3.98$, n.s.			$\chi^2(5)=3.61$; n.s.		
Extrinsic						
Very important	32	28	79:78	4	3	27:22
2	47	50		23	19	
3	16	25	21:33	33	36	69:74
4	5	7		29	28	
5	0	1		7	10	
6	0	0	0:0	4	3	5:5
Not important at all	0	0		1	2	
	$\chi^2(4)=2.50$, n.s.			$\chi^2(6)=2.13$, n.s.		
n	62	113		130	120	

Notes: All values are coded from 1 to 7, with low values describing positive answers and high values describing negative answers. "Ratio" refers to the proportions of responses in the positive, middle and negative areas on the decremental scales (first number) compared to those on the incremental scales (second number).

proportions show a systematic tendency in the postulated direction they tend to support both hypothesis 1 (primacy effect on decremental scales), and hypothesis 2 (recency effect on incremental scales). The most important of these observations is that the postulated differences between scale directions are the same in the Hebrew and the German reading directions, although they are somewhat more pronounced in the former than in the latter.

Comparison of Latent Means

Based on cognitive elaboration theory, we expected differences in means between the incremental and the decremental scales. According to the coding of values, we postulated higher means (i.e., more negative responses) for the incremental scales (hypothesis 3). As already mentioned, the comparison between means of the latent variables *achievement motivation*, *intrinsic job motivation,* and *extrinsic job motivation* assessed by different scale directions was conducted within the German and the Israeli groups. The fact that we used partially different models in the German and the Israeli groups is of minor relevance here because the models for testing the effect of scale direction are equivalent within each group. Measurement equivalence within each group was supported by the data, and differences in latent means express the response-order effect due to decremental and incremental scale direction within the right-to-left (Hebrew) and the left-to-right (German) reading directions.

For the comparison of latent means, we used the incremental scale as a reference group. Table 4 shows that, with one exception, latent means do not differ significantly between scale directions, and that this result holds for the right-to-left (Hebrew) and the left-to-right (German) reading directions. The exception is achievement motivation, where a significant difference between scale directions occurs in the case of the Israeli group. The negative signs for all comparisons reveal the same pattern as that already observed for the proportions: Compared to incremental scales, decremental scales elicited more positive responses. Once again, results are in the postulated direction, but they are mostly not significant. However, because these results are in line with the literature according to which response-order effects on horizontal rating scales are chronically small, we interpret these systematically occurring differences as support for hypothesis 3.

Summary and Discussion

Although the effects of response order on response behavior have been the subject of extensive investigation, these effects have not hitherto been investigated in different reading directions. The results of the present study investigating response-

Table 4 Latent mean differences between incremental and decremental response scales within the right-to-left (Hebrew) and the left-to-right (German) reading directions

	Est.	S.E.	C.R.	p-value
*Reading: Right to Left (Hebrew)**				
Achievement motivation	-0.413	0.174	-2.383	0.017
Job motivation (intrinsic)	-0.021	0.106	-0.193	0.847
Job motivation (extrinsic)	-0.167	0.157	-1.061	0.289
*Reading: Left to Right (German)***				
Achievement motivation	-0.163	0.110	-1.484	0.138
Job motivation (intrinsic)	-0.081	0.099	-0.818	0.413
Job motivation (extrinsic)	-0.106	0.127	-0.813	0.406

Notes: * Model with one cross-loading. ** Model with two residual covariances.

Values coded from 1 to 7 (*applies to me completely* to *does not apply to me at all* and *very important* to *not important at all*, respectively). The reference group is the incremental scale. Est: estimated difference of the latent mean on the decremental compared to the incremental scale. A negative sign indicates more positive (i.e., lower) values on the decremental scale. S.E.: standard error of the estimate. C.R.: critical ratio of the difference. p-value: significance level of the difference.

order effects elicited by scale direction within the right-to-left (Hebrew) and the left-to-right (German) reading directions reveal first and foremost the existence of response-order effects within both reading directions. These effects are of comparable size.

However, the postulated response-order effects are significant only for achievement motivation, which refers to individual self-descriptions and is measured using a question format that is structurally equivalent to an agree/disagree scale. Empirical evidence reveals that these scales are especially susceptible to response-order effects (Liu, Lee, & Conrad, 2015). By contrast, extrinsic and intrinsic job motivation were assessed on a scale that measured the importance ascribed to job characteristics. Because the latter method (known as item-specific question format) implies a more direct manner of question asking (Höhne & Krebs, 2017), it is less susceptible to response-order effects (Saris, Revilla, Krosnick, & Schaeffer, 2010). Hence, the results of the present study add to these findings by revealing that response-order effects that occur in different question formats in the left-to-right (German) reading direction occur similarly in the right-to-left (Hebrew) reading direction.

The differing results with regard to response-order effects for achievement motivation and (extrinsic and intrinsic) job motivation may be due to the specific content of the motivational dimensions. This is especially true for extrinsic job motivation, where indicators address commonly desirable job characteristics such as income and career prospects. Empirical evidence for the apparent immunity of extrinsic job motivation to response-order effects was found by Krebs and Hoffmeyer-Zlotnik (2010). Their interpretation is strongly related to the "hierarchy of importance" described by Toepoel and Dillman (2011), whereby question content takes precedence over scale direction and question format. This implies that a question's content might not be susceptible to response-order effects, irrespective of scale direction and question format. However, this is merely an attempt at an explanation, and it lacks empirical evidence. Furthermore, according to empirical findings regarding intrinsic job motivation, this hierarchy does not apply. Therefore, to learn more about the relation between question content and question format and/or scale direction, we hope that future research will investigate the hierarchical order between question content and different question design strategies. This is especially desirable because the results of the present study reveal that the postulated "hierarchy of importance" seems to exist in the same manner for the left-to-right (German) and the right-to-left (Hebrew) reading directions.

A further promising result of this study is that measurement equivalence was established for decremental and incremental response scales in both reading directions. This finding contributes to knowledge about scale construction in cross-cultural comparison. Especially with respect to this circumstance, further research with different question content would be desirable and necessary.

This study has two limitations. First, our results are based on students' responses, and we therefore have a relatively unique sample. However, this does not fundamentally limit the validity of the empirical findings. The students participated voluntarily and without incentives. Regarding respondents' educational level, the hypotheses were tested under strict conditions. In a general population study with respondents of different ages and educational levels, one could expect that the observed differences between decremental and incremental scales would be more pronounced.

This leads to the second limitation of our study, namely the (mostly) nonsignificant results. Although the general tendency of the results is consistent with, and reinforces, the predictions of cognitive elaboration theory and satisficing theory, the lack of significance is disappointing. On the one hand, it can be attributed to the small case number. Therefore, further research is desirable to investigate response-order effects of decremental and incremental response scales in different reading directions with general population samples. On the other hand, compared to categorical scales, response-order effects for rating scales are chronically small (Sudman et al., 1996), especially when scales are presented horizontally (Höhne

& Lenzner, 2015). Therefore, the investigation of the size of response-order effects in vertical scales in different reading directions would be an appealing topic for further research.

Irrespective of these limitations, this study contributes to existing research and theory by corroborating empirical findings and theoretical reasoning. The similarity of response-order effects in the right-to-left (Hebrew) and the left-to-right (German) reading directions points to the importance of scale direction effects across cultural contexts. The results imply that response-order effects postulated for the left-to-right reading direction are replicated in the right-to-left (Hebrew) reading direction. Considering the differentiation between lack of generalizability (student sample) and failure of replication, our study contributes to this methodological aspect of cross-cultural survey research by showing that response-order effects can be replicated in a different (right-to-left) reading direction.

References

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods and Research, 25,* 381-341.

Alwin, D. F. (2007). *Margins of Error. A Study of Reliability in Survey Measurement.* Hoboken, NJ: Wiley & Sons.

Bishop, G. and Smith, A. (2001). Response-order effects and the early Gallup split-ballots. *Public Opinion Quarterly, 65*(4)*,* 479-505.

Bruner II, G.C. (2013). *Marketing Scales Handbook: The Top 20 Multi-Item Measures Uses in Consumer Research*. Fort Worth, Texas: GCBII Productions.

Byrne, B.M. (2012). *Structural Equation Modeling with Mplus. Basic Concepts, Applications and Programming.* New York, NY: Routledge.

Fowler, F. (1995). *Improving Survey Questions. Design and Evaluation.* Thousand Oaks; CA: Sage.

Fowler, F. Jr., & Cosenza, C. (2008). Writing Effective Questions. In: de Leeuw, E. D.; Hox, J. J. and Dillman, D. A. (Eds.), *International Handbook of Survey Methodology*, 136-160. New York: Lawrence Erlbaum Associates.

Frieze, I. H. (2010) *Cross Cultural Survey of Work and Gender Attitudes 1991-2010.* Retrieved from https://sites.google.com/site/friezewebsite/cross-cultural-survey-of-work-and-gender-attitudes

Höhne, J.K., & Lenzner, T. (2015). Investigating response order effects in web surveys using eye tracking. *Psihologia, 48*(4), 361-377.

Höhne, J. K., & Krebs, D. (2017). Scale direction effects in agree/disagree and item-specific questions: A comparison of question formats. *International Journal of Social Research Methodology.* Published online on May 8, 2017. doi: 10.1080/13645579.2017.1325566

Kieruj, N. D., & Moors, G. (2010). Variations in Response Style Behavior By Response Scale Format in Attitude Research. *International Journal of Public Opinion Research 22*(3), 320-342.

Krebs, D., Berger, M., & Ferligoj, A. (2000). Approaching achievement motivation. Comparing factor analysis and cluster analysis. *New Approaches in Statistical Applications: Metodoloski Zvezki, 16*, 147–171.

Krebs, D., & Hoffmeyer-Zlotnik, J. H. (2010). Positive first or negative first? Effects of the order of answering categories on response behavior. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*, 118–127.

Krebs, D. (2012). The impact of response format on attitude measurement. In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, theories, and empirical applications in the social sciences. Festschrift for Peter Schmidt* (pp. 105–113). Wiesbaden: Springer VS.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measurement in surveys. *Applied Cognitive Psychology, 5*, 213–236.

Krosnick, J. A. (1999). Survey research. *American Review of Psychology, 50*, 537–567.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*, 201–219.

Krosnick, J.A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin, (Eds.), *Survey measurement and process quality* (141–164). New York: Wiley.

Krosnick, J. A.; Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. In M. T. Braverman and J. K. Slater (Eds.), *Advances in Survey Research* (29–44). San Francisco: Jossey-Bass.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In: P. V. Marsden and J. D. Wright (Eds.). *Handbook of survey research*, 2nd ed. (263–312.). Bingley: Emerald Group Publishing Limited.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 5–53.

Liu, M., Lee, S., & Conrad, F.G. (2015). Comparing extreme response styles between agree-disagree and item-specific scales. *Public Opinion Quarterly, 79,* 952–975.

Malhotra, N. (2008). Completion Time and Response Order Effects in Web Surveys. *Public Opinion Quarterly, 72,* 914–934.

Mathews, C.O. (1929). The effect of the order of printed response words on an interest questionnaire. *Journal of Educational Psychology, 30,* 128–134.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review. 63*, 81–97.

Menold, N., & Bogner, K. (2014). Gestaltung von Ratingskalen in Fragebögen. *SDM Survey Guidelines*. GESIS. Mannheim.

Muthén, L. K., & Muthén, B. O. (1998–2010). Mplus user's guide. 6th ed. Los Angeles, CA: Muthén and Muthén.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica,* 104, 1–15.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372–422.

Revilla, M. A. 2013. Measurement invariance and quality of composite scores in a face-to-face and a web survey. *Survey Research Methods*, *7*(1), 17–28.

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods and Research*, *43* (1), 73–97.

Robinson, J. P., Shaver, P. R., & Wrightsman L. S. (Eds.). (1999). *Measures of political attitudes.* Vol. 2 of *Measures of social psychological attitudes* series. San Diego, London: Academic Press.

Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, NJ: Wiley & Sons.

Saris, W.E.; Revilla, M.; Krosnick, J.A., & Schaeffer, E.M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods, 4* (1), 51–79.

Schwarz, N.; Hippler, H. J., & Noelle-Neumann, E. (1992). A cognitive model of response-order effects in survey measurement. In N. Schwarz and S. Sudman (Eds.), *Context effects in social and psychological measurement.* (187–201). New York: Springer.

Schwarz, N.; Knäuper, B.; Hippler, H. J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), (570–582).

Simon, H. A. (1959). Theories of decision making in economics and behavioural science. *American Economic Review, 49*(3), 253–283.

Spence, J.T., & Helmreich, R.L. (1983). Achievement-related motives and behavior. In J.T. Spence (Ed.), *Achievement and Achievement Motives: Psychological and Sociological Approaches* (pp. 10–74). San Francisco, CA: Freeman.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers. The application of cognitive processes to survey methodology.* San Francisco, CA: Jossey-Bass Publishers.

Toepoel, V., & Dillman, D.A. (2011). Words, numbers, and visual heuristics in web surveys: Is there a hierarchy of importance? Social Science Computer Review, 29 (2), 193-207. First published on May 18, 2010. doi: 10.1177/0894439310370070

Toepoel, V., Das, M., & van Soest, A. (2009). Design of web questionnaires: The effect of layout in rating scales. *Journal of Official Statistics, 25*, 509–528.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Weijters, B. Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: *The number of response categories and response category labels.* University of Gent, Fakulteit Economie En Bedrijfskunde, Working Paper, January 2010.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*, 956–972.

Yan, T., & Keusch, F. (2015). The effects of the direction of rating scales on survey responses in a telephone survey. *Public Opinion Quarterly, 79*(1), 145-165.

Appendix

English translation of the German questions (decremental scale direction).

I enjoy being in competition with other people. (Achievement)
It is important to me to perform better than others on a task. (Achievement)
No matter what the activity is, I enjoy being better than others. Achievement)
I try harder when I am in competition with other people. (Achievement)
Applies to me completely – Does not apply to me at all

How important to you is a job ...
... where you can decide for yourself how the work should be done? (Intrinsic)
... that allows you to use your skills and talents? (Intrinsic)
... where you have responsibility for specific tasks? (Intrinsic)
... that allows you to realize your own ideas? (Intrinsic)
... with a high income? (Extrinsic)
... with good promotion prospects? (Extrinsic)
... with clear career prospects? (Extrinsic)
... where you are respected by your superiors? (Extrinsic)
Very important – Not important at all

methods, data, analyses | Vol. 12(1), 2018, pp. 127-150 DOI: 10.12758/mda.2017.16

Exploring Language Effects in Cross-cultural Survey Research: Does the Language of Administration Affect Answers About Politics?

Diana Zavala-Rojas
Universitat Pompeu Fabra

Abstract

We study if the language of administration of a survey has an effect on the answers of bilingual respondents to questions measuring political dimensions. This is done in two steps. In the first we test whether the measurement instruments are equivalent for the same individual in two languages. After measurement invariance is established, we test if latent mean differences are significant across the two languages. We also test if the correlation of the same concept in two languages is equal to one or not. Results show evidence for language effects, the latent correlation is below one, although mean differences are not significant. We use data of the LISS migration panel in a within subject design, respondents answer a questionnaire twice first in Dutch and then in their (second) native language among Arabic, English, German, Papiamento and Turkish.

Keywords: language effects, bilingualism, measurement equivalence

Introduction

Target populations studied in large scale cross-national survey projects are linguistically diverse. In survey projects such as the European Social Survey, and the Survey of Health, Ageing and Retirement in Europe it is a common practice to translate questionnaires when at least 5% of the population is native speaker of a language (Dorer, 2012; SHARE, 2014), but little is known about the consequences and rationale behind this decision (Andreenkova, forthcoming 2018). In the present research we study if the language of administration of a survey influences the answers of *bilingual* respondents to questions measuring *political dimensions*. We define bilingual individuals in terms of language use, that is, individuals who have the ability to write, to speak, to read and to listen in two languages. Furthermore, they use both languages in their daily life: in their main activities such as work or school and with their friends and relatives (Grosjean, 2014).

Language effects in comparative survey research can have different forms; for instance, problematic translations can fail to reproduce the same stimuli across languages (Pennell et al., 2010; Davidov & De Beuckelaer, 2010), or the language of an interview usually activates cultural orientations driving individuals' responses (Luna, Ringberg & Peracchio, 2008; Peytcheva, forthcoming 2018). Language is a strong cultural carrier (Cohen, 2009) and bilingual individuals tend to live in mixed cultural environments. Cultural orientations may influence thoughts, cognitions and behaviour (Oyserman & Lee, 2008), and this in turn may affect the way respondents interpret and answer survey questions. Although translation issues have gained importance in comparative survey methodology, so far the effects of the language of administration on the responses to a questionnaire have received little attention in the field of survey research (exceptions are Peytcheva, forthcoming 2018; Elliot et al., 2012).

Research about language effects in the answers bilingual individuals give to measurement instruments has been conducted mainly in the fields of sociocultural psychology and psycholinguistics. In these two disciplines, even though diverse in methods and approaches, it has consistently been found that the language of administration of a questionnaire has an effect on the answers bilingual individuals give to cultural and self-identity items by activating specific cultural orientations linked to the language of the questionnaire (Chen & Bond, 2010, for a review; Chen, Benet-Martínez, & Ng, 2014). As the proportion of bilingual individuals is different across countries, the potential impact of this effect in cross-national survey research is unknown.

Direct correspondence to
Diana Zavala-Rojas, Universitat Pompeu Fabra (Spain)
E-mail: diana.zavala@upf.edu

To fill this gap, we have carried out a research project in which we test for language effects in a within-subject study of bilinguals in the Netherlands, a country with high linguistic diversity. Participants answered a questionnaire in Dutch and in their (other) native tongue: Arabic, English, German, Papiamento or Turkish. The first step is to test for measurement equivalence. Once equivalence is established, we test whether the correlation of a concept in two languages is equal to one. Third, we test if differences in latent means across languages were significant. The article proceeds as follows: In the next section, we introduce the mechanisms behind the effects of the language of administration on the answers to measurement instruments. Then, we introduce the operationalization of the concepts 'Trust and need of change in institutions' and 'Satisfaction and need of change in politics and the economy' and the models used to test for language effects. Afterwards, we explain the methodology we employ, that is, the procedures regarding the estimation and testing of the models. Next, we present the survey data we use. Finally we summarize the results and discuss the general findings.

Language Effects in Responses to Measurement Instruments

The mechanism behind the adaptation of responses as a function of the language in an interview can be explained by the theoretical frameworks of *acculturation* (Schwartz et al., 2014)with fully bilingual Hispanic participants from the Miami area, to investigate 2 sets of research questions. First, we sought to ascertain the extent to which measures of acculturation (Hispanic and U.S. practices, values, and identifications and *cultural frame switching* (*CFS*)*,* Honget al., 2000). As language is a strong cultural carrier (Cohen, 2009), individuals who master two languages may start an acculturation process, developing into a bicultural person (Grosjean, 2014) by internalizing to some extent the cultural attitudes and values attributable to the second language (Bond & Yang, 1982). Acculturation operates in three dimensions. The first is at the level of social behaviours or *practices*, such as cuisine preferences, language use and the choice of friends. The second is the acquisition of cultural *values*, for instance the importance of individualism versus collectivism. The third dimension is about identification: the attachment to a cultural, ethnic or national group (Schwartz et al., 2010).

CFS takes place when a person uses one system of cultural orientations instead of the other to react to specific social cognitions. This happens when cultural orientations are activated and become highly accessible in the mind of the person. Research has shown that the language of the interview can be a powerful activator of culture-specific mindsets in bilinguals, and individuals' answers to a questionnaire are adjusted accordingly (Bond & Yang, 1982; Chen, Benet-Martínez, & Ng,

2014; Chen & Bond, 2010; Luna et al., 2008; Schwartz et al., 2010, 2014; Yang & Bond 1980).

Previous research about language effects in bilingual individuals has been conducted in most cases with Asian subjects comparing their responses in Chinese and English languages, followed by research on the differences between Spanish and English in Hispanic communities in the United States. However, the dichotomies Chinese-Westerner or Hispanic-Westerner (where Western means English language or American culture) may be very specific cases. Both Chinese and Hispanic cultures emphasize collectivism as an archetypal trait, whereas preference for individualism is regarded as a Western archetype (Yoon, 2010). Respondents from highly communitarian cultures are more sensitized to contextual clues. They may assume that a certain type of culturally oriented response is expected (Lechuga, 2008). Moreover, the distance between Asian cultures and Western culture is perceived as very large (Minkov, 2007).

When language effects have been tested in other cultural contexts, findings have not been replicated completely. It remains unanswered to what extent language effects can be generalized to individuals of cultural backgrounds that are not Chinese or Hispanic. Other languages have been explored in fewer cases: for instance Arabic-French and Arabic-English (Botha, 1968), Afrikaans-English (Botha, 1970), Cebuano (Watkins & Gerong, 1999), French-English (Candell & Hulin, 1986), Greek-English (Richard & Toffoli, 2009; Triandis et al., 1965), Korean-English (Perunovic et al., 2007) and Russian-English (Marian & Neisser, 2000) and, to our knowledge, only one large scale study was conducted in more than 20 languages versus English (Harzing, 2006).

Language effects have been found consistently in responses to questionnaires about cultural dimensions (Benet-Martínez, Lee, & Leu, 2006; Bond & Yang, 1982; Harzing, 2005; Lechuga, 2008; Schwartz et al., 2014; Toffoli & Laroche, 2002; Triandis et al., 1965; Yang & Bond, 1980), personality perceptions (Chen et al., 2014; Chen & Bond, 2010; Ramírez-Esparza et al., 2006), feelings (Marian & Kaushanskaya, 2004; Perunovic et al., 2007), autobiographical memory (Marian & Neisser, 2000; Schrauf & Rubin, 2000), subjective evaluative ratings (Bond, 1985; Elliott et al., 2012; Pierson & Bond, 1982; Toffoli & Laroche, 2002) and self-relevant identity constructs (Dixon, 2007; Kemmelmeier & Cheng, 2004; Pierson & Bond,1982; Ross et al., 2002; Trafimow et al., 1997).

Luna et al., (2008) state that CFS only happens in bicultural bilinguals. The feelings and knowledge that monocultural bilinguals have associated to their second language does not affect how they see themselves. Consistently, several studies have found that language effects are mediated by individual characteristics related to *biculturalism*. Examples are the time in a lifespan and length of exposition to cultural practices of both cultures, and the extent they are perceived as compatible or oppositional; or the *language acquisition*: for instance in which setting the

languages were learned or the time of first exposition to each language (Benet-Martinez et al., 2002; Benet-Martínez & Haritatos, 2005; Dixon, 2007; Ji, Zhang, & Nisbett, 2004; Ross, Xun, & Wilson, 2002).

Benet-Martinez et. al., (2006) found out that biculturals' thinking about culture is more sophisticated than that of monocultural individuals. They are more experienced in dealing with cultural information because of their frequent CFS experiences. As a consequence, biculturals would have more complex cultural representations than monoculturals, but they were not expected to have complex representations in culturally neutral domains, such as geometric figures or landscapes. However with the exception of physical and mental health for which language effects did not emerge (Elliott et al., 2012; Peytcheva, 2008), *culturally neutral topics* have been tested in a few cases. Language effects have been studied in laboratory-settings on culturally neutral topics, being far too neutral, and of no relevance to social or political dimensions. Peytcheva (2018) argues that language effects would likely be present when the cultural specifics evoked by the language prompt cues of to what types of responses are socially accepted. Therefore, in the same survey interview, some items can be affected by language effects while for others this may not be the case.

There are several methodological limitations of most published research. The first is that language effects are tested by mean differences in composite scores of observed variables implicitly assuming that the measures are statistically equivalent across linguistic groups. Measurement equivalence is a prerequisite for cross-cultural comparison of models, relationships and means (Davidov et al., 2014; Meredith, 1993; Vandenberg & Lance, 2000). Before interpreting differences in responses, it is essential to test if the same measurement model on the relationship between indicators and latent variables holds in both languages. Only in few exceptions, measurement equivalence has been established prior to test for language effects in bilingual individuals (Candell & Hulin, 1986; Richard & Toffoli, 2009; Schwartz et al., 2014 test for measurement invariance and language effects).

A second methodological limitation in the analysis of language effects is that manifest variables are not measurement-error free. When differences in observed means have not been found to be significant, the conclusion has been that language effects are negligible. Only when full invariance is found, composite scores can be used directly. When partial invariance is found (Byrne, Shavelson, & Muthén, 1989), latent means should be used, composite scores are not adequate (Saris & Gallhofer, 2014, ch. 16).

A third limitation is that when mean scores are compared, it is, in general, not tested if the conceptual associations that individuals retrieve when they use one language or the other are the same, for instance when testing the strength of the correlation between a latent concept in one and the other language. Richard and Toffoli (2009) found that although the factorial structure of a construct (con-

figural invariance) and the way respondents answered (factor loadings invariance) were the same in Greek and English, the covariances between the latent variables were significantly different across languages. They argued that respondents had different conceptual associations in each language. A test where latent (or observed) mean differences are not significant does not rule out the possibility of language effects. It indicates that the distribution of the variable in the two languages is the same (equality in the location parameter) but that respondents can still have different conceptual associations in each language. In fact, evidence suggests that bilinguals may use different conceptual associations in each language, even in the cases where a literal translation exists (Ji et al., 2004; Luna et al., 2008). For instance, the language of an interview has been found to be a powerful activator of memories, individuals may retrieve auto-biographical experiences associated to the use of one language in consistency with the language of the interview. Marian and Neisser (2000) show that respondents interviewed in Russian (resp. English) remembered more experiences of their Russian-speaking (resp. English-speaking) period of their lives, depending on the language of the interview. For Hispanic bilinguals, autobiographical memories were encoded and retrieved in Spanish for events associated to the use of Spanish language, and in English for events in which English language was used (Schrauf & Rubin, 2000).

In our study we use a different approach to test for language effects. We use a specific application of a LISREL model (Jöreskog & Van Thillo, 1973), which we call in the following sections the *baseline model*. With this model, we test if the relationship across indicators and latent variables is the same in both languages. This is a test for measurement equivalence. Once it is established that the measurement model is equivalent, we are able to test structural relationships of latent variables in two languages. We test if two latent variables represent the same variable of interest by testing if its correlation is equal to one (Jöreskog, 1971; Saris, 1982a, 1982b). In other words, if two latent variables representing the same concepts in different languages had a very high correlation, the variables would be very similar across languages, nevertheless they would have a unique component indicating that they are not exactly the same.

Constructs, Survey Measures, and Models to be Tested

We test two concepts: "Political satisfaction" and "trust in institutions", both having a long tradition in political science and survey research. For these concepts we use a similar operationalization previously used in the European Social Survey Round 7 (European Social Survey 2015). Cultural orientations are known to affect political constructs (Inglehart, 1997; Crothers & Lockhart, 2000); thus, if the language of

the interview activates cultural orientations, bilingual individuals may score differently depending on the language of the interview.

In addition, we develop a measure of respondents‘ perception of political change. We operationalize the concept ‚political change‘ in a survey questionnaire following the three step procedure to formulate survey questions suggested by Saris and Gallhofer (2014). Appendix 1 shows the survey questions used in Model 1 and Model 2.

The first model we test is: ‚Trust and need of change in institutions‘ (Figure 1), consisting of two latent concepts. The first labelled ‚Trust‘ in Dutch institutions reflects the measures ‚trust in the parliament‘, ‚trust in the political parties‘, and ‚trust in the police‘. The second concept, ‚Need of change‘ reflects measures representing evaluative beliefs about the need of change in the way the aforementioned institutions work. Similarly, ‚Satisfaction and need of change in politics and the economy' (Figure 2) includes two concepts: ‚Satisfaction with politics‘ reflecting the indicators for ‚satisfaction with the economy‘, ‚satisfaction with the government‘, and ‚satisfaction with the democracy in the Netherlands‘. The concept ‚Need of change‘ reflects ‚the need of change in the economy‘, ‚the need of change in the way democracy works‘ and, ‚the need of change in the government‘. The left hand side of the figures represents the answers of the Dutch questionnaire, in the right hand side, the model corresponds to the same individuals answering in a second language (among Arabic, English, German, Papiamento and Turkish).

The η_j represent the jth latent variable; the y_{ij} is the ith observed variable for the jth latent trait and ε_{ij} are the disturbance terms; the λ_{ij} are the loadings; τ_{ij} are the intercepts and κ_j the latent means. It is assumed that the disturbance terms have a mean of zero and that they are uncorrelated with the latent variables. The disturbance terms are a combination of random errors and unique components. Thus, the unique components are correlated for the same observed variable in different languages denoted by $cov(\varepsilon_{11}, \varepsilon_{13})$, $cov(\varepsilon_{21}, \varepsilon_{23})$, ..., $cov(\varepsilon_{52}, \varepsilon_{54})$, $cov(\varepsilon_{62}, \varepsilon_{64})$. The other disturbance terms are assumed to be uncorrelated. The latent variables (η_j) are correlated with each other. In order to assign a scale to them, the loading of one observed variable is fixed to one, and the respective intercepts to zero (depicted with a dotted line in the pictures).

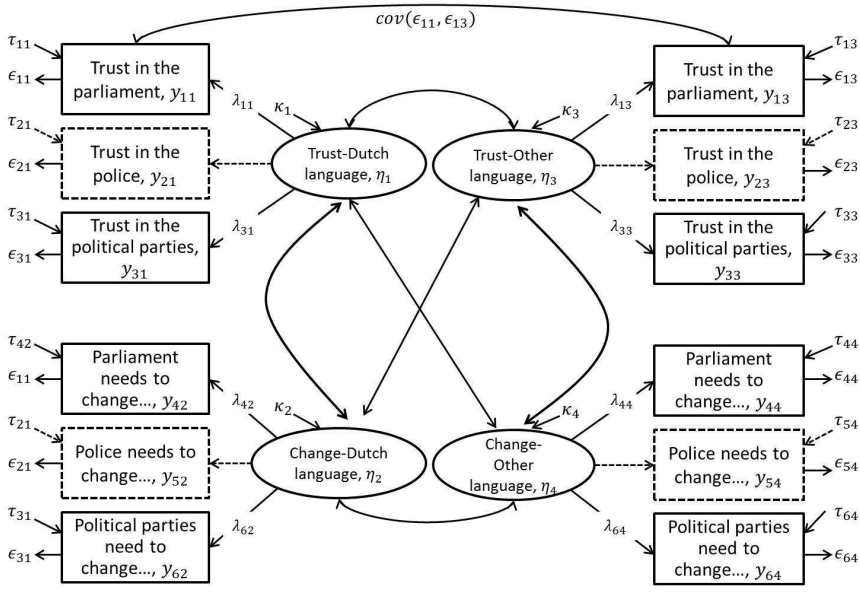

Figure 1 Model 1: Trust and need of change in institutions

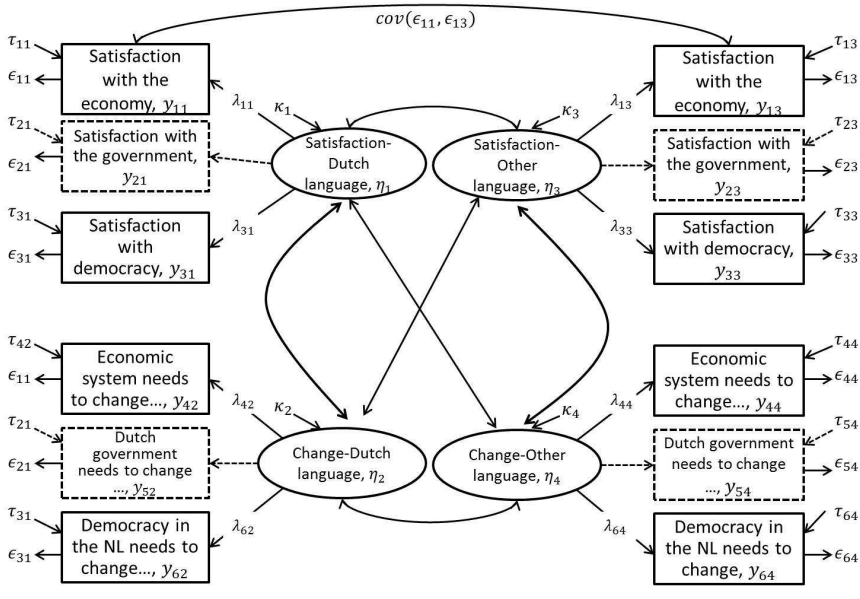

Figure 2 Model 2: ‚Satisfaction and need of change in politics and the economy

Method

We test for the measurement equivalence of measures answered in two languages by fitting a series of models starting with the baseline models shown in Figure 1 and Figure 2, and introducing consecutively equality constraints in the parameters (Davidov et al., 2014; Meredith, 1993; Vandenberg & Lance, 2000)[1]. First, we test that the same configuration of the factorial structure held in both languages. Second, the configural model is restricted to one where the factor loadings are constrained to be equal for the same manifest variable in a different language ($\lambda_{11} = \lambda_{13}$; $\lambda_{31} = \lambda_{33}$; $\lambda_{42} = \lambda_{44}$; $\lambda_{62} = \lambda_{64}$). When this restriction is not rejected, it is implied that comparisons of unstandardized relationships of observed variables across languages can be done. Thirdly, in addition to equivalence in the factor loadings, the intercepts are constrained to be equal ($\tau_{11} = \tau_{13}$; $\tau_{31} = \tau_{33}$; $\tau_{42} = \tau_{44}$; $\tau_{62} = \tau_{64}$). When the restriction in the intercepts is not rejected, it is implied that comparisons of means can also be done across languages.

Once equivalence in the measurement parameters is established, we further constrain the models to test first, whether the correlation between a construct in Dutch and in another language is equal to one ($\rho(\eta_1,\eta_3) = 1$; $\rho(\eta_2,\eta_4) = 1$). Failing this test is interpreted in the sense that the variables „reflect[ing] differences in conceptual associations among the true scores“ (Vandenberg, 2002, p. 142)and that it should be, Vandenberg and Lance elaborated on the importance of conducting tests of measurement invariance and proposed an integrative paradigm for conducting sequences of measurement invariance tests. Building on their platform, the current article addresses some of the shortcomings in our understanding of the analytical procedures. In particular, it points out the need to address (a and that they are not exactly the same, because they have a unique component in each language (Saris, 1982a). To estimate latent correlations and test whether or not they were one, two additional restrictions need to be imposed to the scalar models: the first is to fix the variances of the latent variables to one. The second, fixing the latent covariances of the same concepts in different languages to one. Using these constraints, the model estimates the matrix of standardized latent covariances, which are the latent correlations. Finally, we also test for invariance in the factor means ($\kappa_1 = \kappa_3$; $\kappa_2 = \kappa_4$). This restriction tests for differences between the two languages in the mean latent scores.

1 We estimated the models using Maximum likelihood estimation with ‘lavaan’ package for structural equation modeling (Rosseel 2012) in R3.1.2 statistical environment (R Core Team 2015). All reproducible scripts and the data for this article can be obtained from the author upon request.

Estimation and Testing of the Models

Goodness of fit (GoF) indices of structural equation modelling (SEM) are controversial (Cheung and Rensvold 2002). Commonly used fit criteria such as the Chi-square and the Root Mean Squared Error of Approximation (RMSEA) do not control for Type II error. We use the likelihood ratio test (LRT) in combination with the Judgement Rule (JRule) approach to test our models (Saris, Satorra, and Van der Veld 2009)[2]. The difference in the LRT indicates if the GoF is significantly worse for progressively more restrictive models. The JRule approach (Saris et al. 2009) identifies if fixed or constrained parameters are misspecified. A misspecification occurs if at each level of the equivalence tests specified in the previous section, a parameter has been given a fixed or constrained value, which is incorrect in the population of study (Hu and Bentler 1998). With this approach we can test directly for misspecifications in the models taking into account the power of the test for each misspecification. JRule works by combining knowledge of: (a) the size of the misspecification (expected parameter change); (b) the modification index, its impact on the fit if the parameter was freed in the model; and (c) the power of the test in detecting the misspecification[3]. Only when the modification index is significant and the power of the test low, the parameter is considered misspecified and freed in the models.

Saris et al. (2009) proposed a heuristic approach to choose the threshold for relevant differences. Following this recommendation, we detect standardized loading differences larger than 0.1, and intercept differences larger than 5% of the range of the response scales. As all measures had 11-point scales, this corresponds to intercept differences from 0.55. If a constrained parameter is misspecified according to JRule, it is freed and the null hypothesis of invariance in that restriction rejected. Once measurement equivalence is established, we set a threshold of 0.55 for differences in standardized latent means, which equals 5% of the items' scale. To test for equality of latent covariances/correlations, we restrict them to be equal between groups and test if this restriction was misspecified or not with a threshold of 0.10 for differences. For all decisions, we use a power of the test of 0.80.

Data

We conducted a two wave study between April and June 2013 in the Measurement and Experimentation in the Social Sciences (MESS) Immigrant Panel administered by CentERdata at Tilburg University, The Netherlands. The Immigrant Panel was

2 Appendix 2 reports global fit indexes.

3 The JRule approach for R is available in the 'miPowerFit' function, 'semTools' package (Pornprasertmanit et al. 2014).

a probability based online project in which researchers could submit proposals for fieldwork at no cost. Respondents were recruited based on stratified sampling using the population register as sampling frame. Participants were first and second generations of western and non-western origin of four major migration groups. They were provided with internet and a laptop to answer monthly surveys and received an economic incentive for each completed questionnaire.

The objective of Wave 1 was to select the languages to test for language effects in a within-subject design in Wave 2. The questionnaire included questions in Dutch about language use and knowledge, and questions about politics (see Appendix 1). All participants self-rated their ability in writing, listening, speaking and reading Dutch and their (second) native language in an 11-point scale (from 0 to 10). Wave 1 included 989 bilingual participants. They mentioned 74 languages as their native tongues. We selected the five languages in which respondents had the highest self-reported proficiency and the group was of at least 30 individuals: Arabic, English, German, Papiamento and Turkish. The source questionnaire was developed simultaneously in Dutch and English, translations into the other four languages were done by two independent translators, after which an adjudicator harmonized and decided upon the differences after discussing options with the translators. Questions were pretested with at least one person in each language. We based our procedure on the committee approach proposed by Harkness, Pennell, and Schoua-Glusberg (2004) for survey questionnaires, by involving a team in the translation process, although we simplify it due to budget restrictions.

In the second wave, the questionnaire was presented to 308 bilingual panel members, and it was fully completed by 255 respondents (83%). Due to the small number of individuals per language, the analysis was done within subjects, but it was not possible to separate the different linguistic groups. It was not possible to randomize the order of the languages, therefore, order effects may be present. The results presented in the next section are derived from this final sample size. Table 1 shows the mean and standard deviation of self-reported proficiency in both languages for participants who later on participated in Wave 2 (with the number of respondents by language and completion rates in parenthesis).

Although participants use their (other) native tongue in personal contexts such as at home and with their parents, at school and work, their predominant daily language is Dutch (Table 2). Turkish speakers have a balanced used of both languages with friends, and for German speakers, German language is less frequent in all aspects of life except with their parents.

Wave 2 consisted of three parts. In the first, individuals answered the core questions in Dutch. After that, they answered an unrelated questionnaire about different topics such as ideal body types, nature preservation, and King Willem-Alexander's succession. In the third part, they answered the core questions in Arabic, English, German, Papiamento or Turkish depending on the information they

provided in the first wave. Although memory effects cannot be excluded, they can be controlled for in the case of repetitions in survey interviews by asking other questions in between (Saris and van Meurs 1990).

Table 1 Mean self-reported proficiency in Dutch and target languages (standard deviation)

	Dutch				Target language			
Language group	Write	Read	Speak	Listen	Write	Read	Speak	Listen
English (n=104, 82.5%)	7.6 (2.4)	9.0 (1.4)	8.8 (1.5)	9.0 (1.5)	8.7 (1.7)	9.1 (1.4)	9.1 (1.2)	9.3 (1.3)
Papiamento (n=31, 86.1%)	7.1 (2.7)	8.5 (2.3)	8.6 (1.3)	8.9 (1.3)	6.3 (3.1)	7.4 (2.7)	8.5 (2.2)	8.8 (2.1)
Arabic (n=30, 83.3%)	5.9 (2.4)	7.0 (2.5)	7.0 (2.5)	7.4 (2.4)	7.8 (2.6)	8.2 (2.5)	8.8 (2.1)	9.0 (1.9)
German (n=35, 92.1%)	8.0 (1.8)	9.6 (0.8)	9.2 (1.3)	9.7 (0.7)	7.4 (2.4)	9.1 (1.3)	8.3 (2.1)	9.3 (1.1)
Turkish (n=55, 76.4%)	7.1 (2.5)	8.0 (2.2)	7.8 (2.1)	8.1 (2.0)	7.4 (2.5)	7.3 (2.6)	7.8 (2.2)	8.0 (2.0)

Table 2 Self-reported language use in Dutch and a second language

	Dutch language most frequently used... (%)				Second language most frequently used... (%)			
Language group of:	At work/ school	With friends	At home	With parents	At work/ school	With friends	At home	With parents
Arabic	92.6	56.7	40	0	3.7	33.3	53.3	88.2
English	70.2	81.7	51.9	43	29.8	16.3	47.1	70.7
German	85.7	97.1	85.7	26	8.6	2.9	11.4	47.3
Papiamento	100	70.9	54.7	14.2	--	25.5	45.4	71.2
Turkish	90.2	45.5	21.8	6	7.8	49.1	69.1	88

Note. Percentages adding Dutch and a second language for the same domain do not sum up to 100 when ‚other' language was reported as most used. For example, 56.7% of the Arabic speakers reported Dutch as their most frequently used language with friends, for 33.3% it was Arabic, and for 10% it was another language

Results

Equivalence in the Factorial Structure

Following the JRule test of local misspecifications, the baseline Model 1 (Trust and change in institutions) and Model 2 (Satisfaction and change in politics and the economy) are slightly modified. The p-value of the LRT is significant for the fit of the baseline model versus a model with some correlated errors (Table 3). In Model 1 we introduce two error covariances. The first is between the disturbance terms of the observed variable 'trust in the police' and 'need of change in the way the police works' ($cov(\varepsilon_{21}, \varepsilon_{52}) = cov(\varepsilon_{23}, \varepsilon_{54})$) and the second between 'trust in political parties' and 'need of change in the political parties' ($cov(\varepsilon_{31}, \varepsilon_{62}) = cov(\varepsilon_{33}, \varepsilon_{64})$). Both correlations are constrained to be equal across languages. In Model 2, we introduce three error covariances restricted to be equal between languages: 1) 'satisfaction with the economy' and 'need of change in the economy' $cov(\varepsilon_{11}, \varepsilon_{42}) = cov(\varepsilon_{13}, \varepsilon_{44})$, 'satisfaction with the government' $cov(\varepsilon_{21}, \varepsilon_{52}) = cov(\varepsilon_{23}, \varepsilon_{54})$ and 'need of change in the government' and 'satisfaction with the way democracy works in the NL' and 'change in the way democracy works in the NL' $cov(\varepsilon_{31}, \varepsilon_{33}) = cov(\varepsilon_{62}, \varepsilon_{64})$. Correlated errors improve the fit of the model and they are constrained to be equal across languages. Configural invariance is established because the same linear relationships exist between the indicators and the latent variables in both languages.

Equivalence in the Factor Loadings

Once we establish configural equivalence, we constrain the factor loadings to be equal across languages. As shown in Table 3, the LRT of the configural Model 1 and Model 2 are not significantly different from the restricted models. According to JRule this restriction is not misspecified. Therefore, equivalence in the factor loadings is established in both models.

Equivalence in the Intercepts

There are no significant misspecifications in the restricted intercepts. Furthermore, the LRT does not show that the fit was different between a model constraining loadings and a more restricting one which constrains intercepts. Full measurement equivalence is established in Model 1 and Model 2.

Table 3 Likelihood ratio test - Within subject measurement equivalence in Dutch and a second language

	Model 1					Model 2				
	DF	χ^2	$\Delta\chi^2$	ΔDF	$P(>)\chi^2$	DF	χ^2	$\Delta\chi^2$	ΔDF	$P(>)\chi^2$
Baseline model	42	209.9				42	232.8			
Baseline model + correlated errors	40	158.9	51.04	2	<0.001	39	172.4	60.42	3	<0.001
Invariance of loadings	44	165.0	6.06	4	0.19	43	175.4	2.996	4	0.558
Invariance of intercepts	48	170.8	5.82	4	0.21	47	179.8	4.415	4	0.353

Within-subject Structural Equivalence in Two Languages

Test for Cross-correlations Equal to One

We test whether the correlations between a latent variable in Dutch and the same latent variable in another language was equal to one, $\rho(\eta_1,\eta_3) = 1$; $\rho(\eta_2,\eta_4) = 1$). This is not the case in either Model 1 or in Model 2. Both the LRT and JRule indicate that this restriction should be rejected (Table 4). In Model 1, the correlation between 'trust' in Dutch and 'trust' in a second language is 0.78 ($\rho(\eta_1,\eta_3)$; and 0.64 between 'change' in Dutch and 'change' in a second language ($\rho(\eta_2,\eta_4)$. In Model 2, the correlation between the construct for 'satisfaction' in Dutch and 'satisfaction' in another language is not equal to one, but significantly lower (0.79) ($\rho(\eta_1,\eta_3)$. In the case of the CP 'change', the correlation between Dutch and a second language is of 0.71 ($\rho(\eta_2,\eta_4)$.

Test for Equal Factor Means

The fit Model 1 and Model 2 restricting latent means is not significantly different from the one restricting intercepts. According to JRule, we do not find misspecifications in the equality constraints of the latent means. In Model 2, the LRT shows that the fit of the model restricting latent means is significantly worse than the one which estimates the means without constraints. However, at the threshold level of 0.55 (5% of an 11-point scale), JRule does not indicate any significant differences in latent mean differences. When relaxing the threshold to detect deviations of 0.15 with a power of 0.80, JRule indicates that the equality constraints $\kappa_1 = \kappa_3$ and $\kappa_2 = \kappa_4$ are misspecified. The unstandardized estimate for the factor mean of 'satisfaction' is of 3.61 (se = 0.13) in Dutch language (κ_1) and 3.87 (se = 0.12) in the second language (κ_3). The unstandardized latent mean of 'change' is 6.98 (se = 0.12) in Dutch (κ_2) and 6.81 (se = 0.12) in the respondents' second language (κ_4). This

Table 4 Likelihood ratio test - Within subject differences in latent means and covariances

	Model 1					Model 2				
	DF	χ^2	$\Delta\chi^2$	ΔDF	$P(>)\chi^2$	DF	χ^2	$\Delta\chi^2$	ΔDF	$P(>)\chi^2$
Invariance of intercepts	48	170.8				47	179.8			
Correlations test	54	417.3	246.54	6	<0.001	54	495.4	315.55	7	<0.001
Latent means test	50	174.5	3.76	2	0.15	49	191	11.15	2	<0.004
Latent means with ‚satisfaction' mean free						48	182.6	2.75	1	0.09

result indicates that the mean scores of the underlying constructs that build Model 1 are significantly different in Dutch and in a second language for the same individual, however the difference is estimated around 1.5%. It is rather smaller than the threshold for mean differences established in Section 3.1.

Discussion and Conclusions

In the present study, we explore the effects of the language of the survey interview on the answers of bilingual respondents. Except for translation issues, the study of language effects on respondents' answers has received little attention in comparative survey methodology. As cross national comparative survey research expands to populations of study that are culturally diverse, measurement instruments are translated into more languages and more sampled individuals are themselves bilingual. This motivated the study of the potential effects that the language of the survey has on bilingual individuals. A limitation of this study is that the sample size is not large enough to divide the analysis by linguistic group in the bilingual sample, so further research is needed on specific cultural groups. A second limitation is that although the survey questions were repeated in the same survey interview, the true score for the same individuals using a different language may change with the passage of time or may include memory effects, thus changes may not only be due to switching to a different language. Nevertheless, this limitation is inherent to within-subject studies. A third limitation is that our findings cannot be generalized to other themes, they only hold for the tests in this study. Therefore, more research is needed to investigate the extent of language effects in different topics asked by the means of survey questionnaires.

Three specific research questions are addressed in the present study. The first is to investigate if language effects would emerge in bilingual individuals of cultural backgrounds different from those tested in the majority of published articles (Asian and Hispanic descendants). In our study, participants are bilinguals with Dutch as their main language. The second question is if language effects would emerge in political constructs: the reason for this research question is that so far, cultural and self-identity constructs have been explored in the literature rather than political topics. The third is to challenge the classical approach of testing for language effects comparing observed means of composite scores by testing whether the correlation of a latent variable in two languages is one.

In a first step we tested for within-subject measurement equivalence to confirm that our measures in two languages are invariant. Testing for measurement equivalence between languages has been seldom performed in past research, and it is a prerequisite for statistical comparison of survey items across cultures, languages and groups (Davidov et al., 2014). In a second step we tested for differences in latent correlations and means.

The first conclusion is that the measures we employ for the concepts in Model 1, 'Trust and need of change in institutions' and in Model 2, 'Satisfaction and need of change in politics and the economy' are statistically equivalent across languages. The second conclusion is that the language in a survey questionnaire affects to some extent the answers of bilingual respondents to political dimensions. We find, in both models, that the correlation between a latent variable measured by the same questions in Dutch and in a different language is not equal to one but significantly lower.

This is relevant to substantive research using these concepts because if the factors in Dutch and in another language have a very high correlation, the impact of each of them on a third variable will be difficult to distinguish. For instance, the larger the correlation between "political satisfaction" in Dutch and in another language, the more similar effect they have on "political participation". However, when the correlation is low, the association of "political satisfaction" with other variables of interest depends to some extent on the language of the survey measures. This would not be a problem if language effects were consistent across topics, but as we summarize in the literature review section, this is not the case.

Borrowing from cultural psychology the theoretical framework of *cultural frame switching* (CFS) (Hong et al., 2000), we interpret our results arguing that respondents made use of different conceptual associations in each language. As each language is associated with language specific cultural orientations, our results indicate that respondents shifted their cultural frame of reference when answering in the different languages.

However, factor mean differences did not emerge. This result indicates that language effects can be present even in the case when significant differences in

latent means do not emerge. Latent mean differences indicate a difference in the location of the parameters of the distribution of the latent variable[4].

Implications for Survey Methodology

Survey questions are measurement instruments of opinions. If the correlation between the same latent variable in two languages is not one, apparently it would follow that for certain topics, bilingual individuals are able to express two opinions, each triggered by cultural associations evoked by the language of the survey. The first implication of our findings for the design of surveys with multilingual samples is that the decision of the interview language should receive a more important role in the design of surveys. Andreenkova (2018) analyzes documentation on language choice in six large comparative survey projects finding out that information was very limited. The author concludes that more research needs to be done to design strategies for language allocation in bilingual populations, considering for instance, language usage and proficiency inquired from to the respondent at the beginning of the interview and using this information to select the language of the main interview. This would require interviewer training but also increasing survey agencies' awareness about the effects of the language of the interview.

Another possibility would be to give respondents two questionnaires in two different languages, as we did in this study, and average their opinion. From an operational point of view this solution is not optimal: For instance, it increases costs, increases cognitive burden on the respondent, increases the length of the interview and introduces potential memory effects. A third option (suggested in Richard & Toffoli, 2009) would be to randomize the questionnaires across languages. In a survey like the one presented in this study that would have meant that a random group of respondents would have answered in Dutch and another group in a second language. Although this option is statistically sound because differences across languages would cancel out, it is not operational in a comparative survey. The linguistic characteristics of the target population and of the individuals in the sampling frame are in general unknown before the data collection. Thus, the size of the random groups would be unknown as well. Moreover, functional bilingualism implies the combined abilities of writing, speaking, reading, and listening in two languages, and it also implies usage of both languages in their daily life (Grosjean, 2014). It does not imply that respondents feel fully comfortable answering certain topics in both languages.

Summing up, given the increasing evidence that language can affect responses to questionnaires in social and political surveys and in psychological instruments,

4 Very small significant latent means were found in Model 2, but they were well below the set threshold to consider them relevant.

providing an optimal solution on the choice of the language of the interview seems to be a clear aspect of comparative survey methodology that should receive more attention.

References

Andreenkova, A. (forthcoming 2018) How to Choose Interview Language in Different Countries. In: Johnson, T.P., Pennell, B.-E., Stoop, I. & Dorer, B. (Eds.). (forthcoming). Advances in Comparative Survey Methods: Multinational, Multiregional and Multicultural Contexts (3MC). Wiley Series in Survey Methodology. New York: John Wiley & Sons.

Benet-Martínez, V., & Haritatos, J. (2005). Bicultural Identity Integration (BII): Components and Psychosocial Antecedents. *Journal of Personality*, *73*(4), 1015–1050. https://doi.org/10.1111/j.1467-6494.2005.00337.x

Benet-Martínez, V., Lee, F., & Leu, J. (2006). Biculturalism and Cognitive Complexity: Expertise in Cultural Representations. *Journal of Cross-Cultural Psychology*, *37*(4), 386–407. https://doi.org/10.1177/0022022106288476

Benet-Martinez, V., Leu, J., Lee, F., & Morris, M. W. (2002). Negotiating Biculturalism: Cultural Frame Switching in Biculturals with Oppositional Versus Compatible Cultural Identities. *Journal of Cross-Cultural Psychology*, *33*(5), 492–516. https://doi.org/10.1177/0022022102033005005

Bond, M. H. (1985). Language as a Carrier of Ethnic Stereotypes in Hong Kong. *The Journal of Social Psychology*, *125*(1), 53–62. https://doi.org/10.1080/00224545.1985.9713508

Bond, M. H., & Yang, K.-S. (1982). Ethnic Affirmation Versus Cross-Cultural Accommodation: The Variable Impact of Questionnaire Language on Chinese Bilinguals from Hong Kong. *Journal of Cross-Cultural Psychology*, *13*(2), 169–185. https://doi.org/10.1177/0022002182013002003

Botha, E. (1968). Verbally Expressed Values of Bilinguals. *The Journal of Social Psychology*, *75*(2), 159–164. https://doi.org/10.1080/00224545.1968.9712488

Botha, E. (1970). The effect of language on values expressed by bilinguals. *Journal of Social Psychology*, *80*(2), 143. Retrieved from http://search.proquest.com/docview/1290717919?accountid=14708

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456.

Candell, G. L., & Hulin, C. L. (1986). Cross-Language and Cross-Cultural Comparisons in Scale Translations: Independent Sources of Information about Item Nonequivalence. *Journal of Cross-Cultural Psychology*, *17*(4), 417–440. https://doi.org/10.1177/0022002186017004003

Chen, S. X., Benet-Martínez, V., & Ng, J. C. K. (2014). Does Language Affect Personality Perception? A Functional Approach to Testing the Whorfian Hypothesis. *Journal of Personality*, *82*(2), 130–143. https://doi.org/10.1111/jopy.12040

Chen, S. X., & Bond, M. H. (2010). Two languages, two personalities? Examining language effects on the expression of personality in a bilingual context. *Personality and Social Psychology Bulletin*, *36*(11), 1514–1528.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cohen, A. B. (2009). Many forms of culture. *American Psychologist*, *64*(3), 194.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75.

Dixon, D. J. (2007). The effects of language priming on independent and interdependent self-construal among Chinese university students currently studying English. *Current Research in Social Psychology*, *13*, 1–9.

Dorer, B. (2012). *Round 6 Translation Guidelines.* Mannheim.

Elliott, M. N., Edwards, W. S., Klein, D. J., & Heller, A. (2012). Differences by Survey Language and Mode among Chinese Respondents to a CAHPS Health Plan Survey. *Public Opinion Quarterly*, *76*(2), 238–264. https://doi.org/10.1093/poq/nfs020

European Social Survey. (2015). ESS Round 7: European Social Survey Round 7 Data. Bergen: Norwegian Social Science Data Services, Norway – Data Archive and distributor of ESS data.

Grosjean, F. (2014). Bicultural bilinguals. *International Journal of Bilingualism*. https://doi.org/10.1177/1367006914526297

Harkness, J. A., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Survey Questionnaire Translation and Assessment. In *Methods for Testing and Evaluating Survey Questionnaires* (pp. 453–473). John Wiley & Sons, Inc. https://doi.org/10.1002/0471654728.ch22

Harzing, A.-W. (2005). Does the Use of English-language Questionnaires in Cross-national Research Obscure National Differences? *International Journal of Cross Cultural Management*, *5*(2), 213–224. https://doi.org/10.1177/1470595805054494

Harzing, A.-W. (2006). Response Styles in Cross-national Survey Research: A 26-country Study. *International Journal of Cross Cultural Management*, *6*(2), 243–266. https://doi.org/10.1177/1470595806066332

Hong, Y., Morris, M. W., Chiu, C., & Benet-Martínez, V. (2000). Multicultural minds: A dynamic constructivist approach to culture and cognition. *American Psychologist*, *55*(7), 709–720. https://doi.org/10.1037/0003-066X.55.7.709

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453.

Ji, L., Zhang, Z., & Nisbett, R. E. (2004). Is It Culture or Is It Language? Examination of Language Effects in Cross-Cultural Research on Categorization. *Journal of Personality and Social Psychology*, *87*(1), 57–65. https://doi.org/10.1037/0022-3514.87.1.57

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. https://doi.org/10.1007/BF02291366

Jöreskog, K. G., & Van Thillo, M. (1973). LISREL. Department of Statistics: University of Uppsala.

Kemmelmeier, M., & Cheng, B. Y.-M. (2004). Language and Self-Construal Priming: A Replication and Extension in a Hong Kong Sample. *Journal of Cross-Cultural Psychology*, *35*(6), 705–712. https://doi.org/10.1177/0022022104270112

Lechuga, J. (2008). Is Acculturation a Dynamic Construct?: The Influence of Method of Priming Culture on Acculturation. *Hispanic Journal of Behavioral Sciences*, *30*(3), 324–339. https://doi.org/10.1177/0739986308319570

Luna, D., Ringberg, T., & Peracchio, L. A. (2008). One Individual, Two Identities: Frame Switching among Biculturals. *Journal of Consumer Research*, *35*(2), 279–293.

Marian, V., & Kaushanskaya, M. (2004). Self-construal and emotion in bicultural bilinguals. *Journal of Memory and Language*, *51*(2), 190–201. https://doi.org/10.1016/j.jml.2004.04.003

Marian, V., & Neisser, U. (2000). Language-Dependent Recall of Autobiographical Memories. *Journal of Experimental Psychology: General*, *129*(3), 361–368. https://doi.org/10.1037/0096-3445.129.3.361

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Minkov, M. (2007). *What makes us different and similar: A new interpretation of the World Values Survey and other cross-cultural data*. Klasika i Stil Publishing House.

Oyserman, D., & Lee, S. W. S. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. *Psychological Bulletin*, *134*(2), 311.

Pennell, B.-E., Harkness, J. A., Levenstein, R., & Quaglia, M. (2010). Challenges in Cross-National Data Collection. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 269–298). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470609927.ch15

Perunovic, E., Wei, Q., Heller, D., & Rafaeli, E. (2007). Within-Person Changes in the Structure of Emotion: The Role of Cultural Identification and Language. *Psychological Science*, *18*(7), 607–613. https://doi.org/10.1111/j.1467-9280.2007.01947.x

Peytcheva, E. (2008). Language of administration as a cause of measurement error. In *AAPOR*. New Orleans.

Pierson, H. D., & Bond, M. H. (1982). How Do Chinese Bilinguals Respond To Variations of Interviewer Language and Ethnicity? *Journal of Language and Social Psychology*, *1*(2), 123–139. https://doi.org/10.1177/0261927X8200100203

R Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.r-project.org/

Ramírez-Esparza, N., Gosling, S. D., Benet-Martínez, V., Potter, J. P., Pennebaker, J. W., & Ramírez-Esparza, J. W. (2006). Do bilinguals have two personalities? A special case of cultural frame switching. *Journal of Research in Personality*, *40*(2), 99–120.

Richard, M.-O., & Toffoli, R. (2009). Language influence in responses to questionnaires by bilingual respondents: A test of the Whorfian hypothesis. *Journal of Business Research*, *62*(10), 987–994. https://doi.org/10.1016/j.jbusres.2008.10.016

Ross, M., Xun, W. Q. E., & Wilson, A. E. (2002). Language and the Bicultural Self. *Personality and Social Psychology Bulletin*, *28*(8), 1040–1050. https://doi.org/10.1177/01461672022811003

Rosseel, Y. (2012). {lavaan}: An {R} Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Saris, W. E. (1982a). Different questions, different variables? In C. Fornell (Ed.), *A second generation of multivariate analysis. 2. Measurement and evaluation* (First, Vol. 2). New York: Praeger Publishers.

Saris, W. E. (1982b). Linear structural relations. In C. Fornell (Ed.), *A second generation of multivariate analysis: Methods* (First, Vol. 1). New York: Praeger Publishers.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Saris, W. E., & van Meurs, A. (1990). Evaluation of measurement instruments by meta-analysis of multitrait multi-method studies. *Proceedings, Amsterdam, 1989, Amsterdam: North Holland, 1990, Edited by Saris, W.E.; Meurs, A.van*, *1*. Retrieved from http://adsabs.harvard.edu/abs/1990emim.conf.....S

Schrauf, R. W., & Rubin, D. C. (2000). Internal languages of retrieval: The bilingual encoding of memories for the personal past. *Memory & Cognition*, *28*(4), 616–623.

Schwartz, S. J., Benet-Martínez, V., Knight, G. P., Unger, J. B., Zamboanga, B. L., Des Rosiers, S. E., … Szapocznik, J. (2014). Effects of language of assessment on the measurement of acculturation: Measurement equivalence and cultural frame switching. *Psychological Assessment*, *26*(1), 100–114. https://doi.org/http://psycnet.apa.org/doi/10.1037/a0034717

Schwartz, S. J., Unger, J. B., Zamboanga, B. L., & Szapocznik, J. (2010). Rethinking the concept of acculturation: Implications for theory and research. *2American Psychologist*, *65*(4), 237–251. https://doi.org/10.1037/a0019330

SHARE. (2014). The Survey of Health, Ageing and Retirement in Europe. Retrieved March 13, 2014, from http://www.share-project.org/

Toffoli, R., & Laroche, M. (2002). Cultural and language effects on Chinese bilinguals' and Canadians' responses to advertising. *International Journal of Advertising*, *21*(4), 505–524. https://doi.org/10.1080/02650487.2002.11104948

Trafimow, D., Silverman, E. S., Fan, R. M.-T., & Fun Law, J. S. (1997). The Effects of Language and Priming on the Relative Accessibility of the Private Self and the Collective Self. *Journal of Cross-Cultural Psychology*, *28*(1), 107–123. https://doi.org/10.1177/0022022197281007

Triandis, H. C., Davis, E. E., Vassiliou, V., & Nassiakou, M. (1965). *Some Methodological Problems Concerning Research Negotiations Between Monoinguals.*

Vandenberg, R. J. (2002). Toward a Further Understanding of and Improvement in Measurement Invariance Methods and Procedures. *Organizational Research Methods*, *5*(2), 139–158. https://doi.org/10.1177/1094428102005002001

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Watkins, D., & Gerong, A. (1999). Language of Response and the Spontaneous Self-Concept: A Test of the Cultural Accommodation Hypothesis. *Journal of Cross-Cultural Psychology*, *30*(1), 115–121. https://doi.org/10.1177/0022022199030001007

Yang, K.-S., & Bond, M. H. (1980). Ethnic Affirmation by Chinese Bilinguals. *Journal of Cross-Cultural Psychology*, *11*(4), 411–425. https://doi.org/10.1177/0022022180114002

Yoon, K.-I. (2010). *Political culture of individualism and collectivism*. The University of Michigan.

Appendix 1

Survey questions administered in both languages

Model 1: Institutions: trust and change

Concept 1: Trust in institutions[5]

We will ask some questions about your level of trust in some institutions, 0 indicates complete distrust and 10 complete trust.

Overall, to what extent do you trust the parliament?
How much do you personally distrust or trust the police?
How much do you personally trust the political parties?

Complete distrust					Neither distrust nor trust					Complete trust
0	1	2	3	4	5	6	7	8	9	10

Concept: Need of change in the institutions

The next questions are about change in institutions, 0 indicates that the institution does not need to change the way it works and 10 indicates that it needs to completely change.

How much do you think that the Dutch parliament needs to change the way it works?

How much you think that the police needs to change the way it works to protect people like you?

To what extent do political parties need to change the way they work?

No need to change at all										Completely
0	1	2	3	4	5	6	7	8	9	10

5 The response scales were shown following each question, not in grids.

Model 2: Politics and the economy: satisfaction and change

Concept 1: Satisfaction with politics and the economy

Now we will ask you some questions about your satisfaction with some aspects of politics and the economy. Use a scale from 0 to 10 where 0 means you are completely dissatisfied and 10 means you are completely satisfied.

How satisfied are you with the present state of the economy in the Netherlands? Overall, how satisfied are you with the way the Dutch government is doing its job? And overall, how satisfied are you with the way democracy works in the Netherlands?

Completely dissatisfied					Neither dissatisfied nor satisfied					Completely satisfied
0	1	2	3	4	5	6	7	8	9	10

Concept 2: Need of change in politics and the economy

We will ask you about the level of change you think some aspects of in politics and the economy need, 0 indicates 'there is no need at all to change' and 10 is that 'it needs to change completely'.

To what extent does the economic system in the Netherlands need to change?

Overall, to what extent does the Dutch government need to change the way it is doing its job?

To what extent does the way democracy work in the Netherlands needs to change?

Not need at all to change										Completely
0	1	2	3	4	5	6	7	8	9	10

Appendix 2

Global fit indexes of the models of models

Model 1. Trust and need of change in institutions

	DF	Chi-square	p-value	RMSEA	90 % confidence interval for RMSEA	CFI	SRMR
Baseline model	42	209.9	0	0.125	0.109, 0.142	0.917	0.071.
Baseline model + correlated errors	40	158.9	0	0. 108	0. 091, 0.126	0.941	0.060.
Factor loadings invariance	44	165	0	0.104	0. 087, 0.121	0.94	0.63.
Invariance of intercepts	48	170.8	0	0.1	0.084, 0.117	0.939	0.064.
Test of latent means differences	50	174.5	0	0.099	0. 083, 0.115	0.938	0.064.
Test of latent correlations = 1	54	417.3	0	0.162	0.148, 0.177	0.82	0.119.

Model 2. Satisfaction and need of change in politics and the economy

Baseline model	42	232.8	0	0.113	0.117, 0.150	0.916	0.072.
Baseline model + correlated errors	39	172.4	0	0. 116	0. 098, 0.134	0.941	0.070.
Factor loadings invariance	43	175.4	0	0.11	0. 093, 0.127	0.942	0.72.
Invariance of intercepts	47	179.8	0	0.105	0.089, 0.122	0.942	0.073.
Test of latent means differences	49	191	0	0.107	0. 091, 0.123	0.938	0.075.
Latent means test after freeing ‚sat' mean	48	182.6	0	0. 105	0.089, 0.121	0.941	0.073.
Test of latent correlations = 1	54	495.4	0	0.179	0.165, 0.194	0.806	0.239.

methods, data, analyses | Vol. 12(1), 2018, pp. 151-176 DOI: 10.12758/mda.2017.15

Education in OECD's PIAAC Study: How Well do Different Harmonized Measures Predict Skills?

Silke L. Schneider
GESIS – Leibniz Institute for the Social Sciences

Abstract

The comparable measurement of educational attainment is a challenge for all comparative surveys and cross-national data analyses. While education is an important predictor or control variable in many research contexts, it is particularly important when studying education and education-related outcomes such as skills or labor market chances. This study evaluates the cross-nationally comparable measurement of education in OECD's Programme for the International Assessment of Adult Competencies, PIAAC, in terms of its construct validity when predicting general basic skills. In order to do so, the predictive power of country-specific (i.e. non-comparable) education variables is compared to the predictive power of different cross-nationally harmonized variables, namely the detailed ISCED-based coding scheme used in PIAAC, ISCED 2011 and 1997 levels, the broad education levels 'low, medium, high', ES-ISCED, as well as years of education. The analyses consist in sets of country-wise linear regressions, taking PIAAC's plausible values and complex sampling into account, and use adjusted R^2 as the indicator for predictive power and validity. The results show that while harmonization into a detailed coding scheme such as the most detailed comparable variable available in PIAAC does not entail large losses of information, the way this variable is further simplified plays a major role for validity. The paper also highlights shortcomings of the detailed variable from a theoretical point of view, such as the lack of differentiation of vocational and general education and other markers of educational content and quality, which are important aspects both for skill development as well as the labor market outcomes of education, and of the country-specific measures of education, which may make the detailed PIAAC education variable look better than it actually is.

Keywords: Measurement; Educational attainment; Skills; Comparative research; Education; Data quality; Survey

Introduction

An important challenge in comparative survey research is how to make data comparable or 'functionally equivalent' across countries (Przeworski & Teune 1970). The underlying process is called 'harmonization' (Wolf et al., 2016; Hoffmeyer-Zlotnik & Wolf, 2003), especially when speaking about the comparability of individual variables (rather than e.g. sampling or fieldwork procedures). Harmonizing survey data cross-nationally entails the risk of 'harmonizing away' meaningful information (Granda et al., 2010). When a harmonized variable carries less information than a non-harmonized one, and the amount of information loss differs across countries, the comparability of the harmonized measure is necessarily limited. This is an important element of comparison error (Smith 2011), a main impediment of successful comparative survey research.

The comparability of background variables such as ethnicity, education or social class (see e.g. Schneider et al., 2016; Braun & Mohler, 2003) has mostly been researched regarding the education variable. This is for two reasons: Firstly, education is a major independent variable in numerous statistical models of survey micro data, either as control or substantive variable, and thus maybe the most important of all background variables (Smith, 1995). Secondly, its harmonization is, because of the stark institutional differences between educational systems, particularly difficult (Braun & Müller, 1997). Cross-national educational attainment levels such as ‚primary education' or ‚tertiary education', even if translated correctly, are likely to be interpreted differently by respondents in different countries depending on features of their educational systems. Therefore, the state of the art for cross-national surveys is to use country-specific questionnaire items to collect information on respondents' educational attainment (Schneider, 2016). The resulting country-specific education variables are then recoded into a cross-national variable after data collection. This approach is called ex-ante output harmonization (Wolf et al., 2016; Ehling, 2003). Today, most surveys use UNESCO's International Standard Classification of Education (ISCED, UNESCO Institute for Statistics, 2012) for harmonizing education variables.

However, there is no agreement on which specific ISCED-based variables to provide to data users – three broad levels, main ISCED levels, or whether sub-categories within levels representing different types of education also need to be taken into account. The method of comparative construct validation is fairly established today for evaluating the comparative validity of harmonized education variables

Direct correspondence to
Silke L. Schneider, GESIS – Leibniz Institute for the Social Sciences,
P.O. Box 122155, 68072 Mannheim, Germany
E-mail: silke.schneider@gesis.org

in cross-national survey data. These analyses consist in sets of country-wise linear regressions, and usually use adjusted R^2 as the indicator for predictive power or validity. Prior research using this method (Schneider, 2010; Kerckhoff & Dylan, 1999; Kerckhoff et al., 2002; Kieffer, 2010; Müller & Klein, 2008; Braun & Müller, 1997) has generally concluded that the education variables in comparative surveys, including those based on ISCED, contain comparison error, especially (but not exclusively) due to the way that country-specific education categories are aggregated into supposedly comparable, broader categories.

This paper adds to this research using the OECD's Programme for the International Assessment of Adult Competencies, PIAAC (OECD, 2013; OECD, 2016a). In addition to not having been the object of a comparative construct validation of the harmonized education variable yet, PIAAC also offers new validation variables that have so far not been exploited for a comparative construct validation, namely literacy and numeracy skills. The relationship between educational attainment and skills is expected to be fairly strong (and thus sensitive to measurement quality) because one important aim of formal education and training systems is skill production (see e.g. Hall & Soskice, 2001). Because of the close relationship between educational attainment and skills, if educational attainment is not well measured, in statistical models using both as independent variables, unmeasured heterogeneity in education may be picked up by the measure of skills (confounding). It is thus of great importance in a survey of adult skills that educational attainment is measured with a high degree of reliability and validity. Such an analysis will also help us better understand the relationship between educational qualifications and skills (Heisig & Solga, 2015).

This paper builds on the work by Schneider (2010), which used occupational status as the validation variable, and evaluates the harmonized educational attainment measures employed in PIAAC. It answers the following research questions:

1. How comparable across countries, in terms of comparative validity, is the most detailed comparative education variable provided in the PIAAC data set?
2. Do we find the same result for the PIAAC data that were previously found for the ESS and other surveys, namely that main education levels and nominal years of education diminish comparative validity? How does ISCED 2011 fare, compared with ISCED 1997?
3. Could a differently aggregated education variable, such as the European Survey Version of ISCED (ES-ISCED) proposed in Schneider (2010), improve the comparative validity of education measures in PIAAC?

This paper starts out by distinguishing dimensions of education and theorizing about their relationship with general basic skills. Then, the PIAAC data and analysis methods will be presented, as well as the harmonized measures of education available in PIAAC. Here the implications of the theoretical rationale for the meas-

urement of educational attainment are also presented. After presenting the empirical results, the paper will summarize and conclude with some practical recommendations for the next Cycle of PIAAC, which will also be relevant to other future cross-national surveys as well as research using existing data.

Dimensions of Education and General Basic Skills

From a theoretical point of view, education and skills are expected to be fairly closely related. In modern societies, formal education is an important source of general basic skill development and 'human capital' (Becker, 1964; OECD, 2013; OECD, 2016a). Examinations in formal education aim to validate the successful acquisition of knowledge, skills and competences, and give legitimacy to subsequently achieved advantageous social positions (Weber, 1922). Consequently, formal educational qualifications are the most common indicator for educational attainment in surveys.

Formal education is not homogeneous but differs in terms of quality, content and type in very complex ways (Smith, 1995). The educational systems in most developed countries provide alternative programs within education levels. Depending on their specific goals and curricula, different types of educational programs can be expected to lead to different levels of general basic skills. In the following, the dimensions of education distinguished by Smith (1995) - quantity, content, quality and type - are examined with respect to their implications for general basic skills, and hypotheses formed for measuring education in such a way as to optimize the prediction of skills by education.[1]

The first dimension of education is *quantity*. From a human capital point of view (Becker, 1964), the longer children go to school, and the higher the level of education eventually reached by youth and young adults, the stronger we expect their literacy and numeracy skills to be. The better an education measure reflects quantity, the better it is thus expected to predict general basic skills (Hypothesis 1).

The second dimension is *content*, i.e. "what is being taught" (Smith, 1995, p.218). Some (especially European) countries track children in lower secondary education already into programs with different content, preparing for different labor market 'careers' (Haller et al., 1985; Braun & Müller, 1997; König et al., 1988). From upper secondary education onwards, *most* (if not all) countries offer different educational programs with specialized content, mostly differentiating university preparatory general education and vocational programs preparing for the labor market. In vocational education, students spend some of their time learning

1 It is important to note that we simplify these dimensions substantially here, compared to the rich array of indicators that Smith himself has to offer for each of them.

practical skills directly relevant to the labor market. In contrast, in general education, most learning time is spent on text and number based tasks. Therefore, at the simplest level, the better an education measure distinguishes between vocational and general programs, the better it is expected to predict general basic skills (Hypothesis 2).

The third dimension of education is *quality*. In countries offering different educational programs or institutional settings at any single level of education, these may differ not only with respect to their curricular content but also their skill (and social) selectivity, an important indicator for the quality of education (Smith, 1995). For example, many countries especially in Eastern Europe have different types of vocational upper secondary education programs (see e.g. Saar, 2008; Straková, 2008; Bukodi et al., 2008). Some of them give access to higher education, while others do not. Typically, those providing access to higher education are more selective and academically demanding, while those only preparing for the labor market are less so. This results in higher skills of graduates from the former programs, which are however typically already evident when *entering* the program and are thus not or only partially (e.g. through the dimension of content, see above) *caused* by the program. A similar argument can be made for tracking in lower secondary education, where different programs may work at different standards. We thus expect education measures that differentiate educational categories by skill selectivity or institutional setting to better predict adult skills than measures not making such a distinction (Hypothesis 3).[2]

The fourth dimension of education according to Smith (1995) is *type*, which consists in several distinct classification systems that partly overlap with those previously discussed. A distinction by type not yet covered but useful here is the place of learning, where we can distinguish entirely school-based programs from programs combining schooling and work, as in apprenticeship programs in mostly German-speaking countries, and on the job training (see e.g. Allmendinger, 1989), where the latter does not count as formal education. Because of the more strongly theoretical content and book-based learning, we can expect the completion of school-based vocational programs to be related to higher general basic skills than apprenticeship programs, where practical learning plays a more prominent role. Therefore, education measures distinguishing between school-based and apprenticeship programs are expected to better predict general basic skills than measures not making such a distinction (Hypothesis 4).

---

2 In many countries, content and quality of education are overlapping dimensions: academically or generally oriented programs are usually more selective and provided in specialized institutional settings (such as the prototypical Gymnasium or traditional university), while vocationally or professionally oriented programs are - at least at the secondary level - less selective and, in countries with differentiated vocational training systems, provided in a variety of institutional settings.

To summarize, a valid comparable measure of educational attainment that well reflects skills may need to differentiate types of formal education in addition to levels of education, ideally in terms of tracks in lower secondary schooling, programme orientation, and, especially within vocational education, selectivity and place of learning. Measures that simplify education by reducing it to one dimension, such as broad levels of education or duration in terms of years of education, can be expected to function less well and less consistently across countries in predicting general basic skills. The dimensions discussed here may also help explain *why* country-specific measures sometimes do a better job at predicting general basic skills than comparative measures.

Data and Methods

The Programme for the International Assessment of Adult Competencies

OECD's Programme for the International Assessment of Adult Competencies (PIAAC) is a cross-national large-scale survey assessing the general basic skills of the adult population – literacy, numeracy and problem solving in technology-rich environments – that are considered essential for successful participation in today's societies (OECD, 2016a). While skills are directly assessed using psychological tests, information on demographic characteristics, education, labor market participation and other indicators are collected using a background questionnaire. Data for the first set of countries (round 1, 24 countries[3]) were collected in 2011/2012, and for a second set of countries (round 2, 9 countries[4]) in 2014/2015. The target population consisted of individuals aged 16 to 65. Multi-stage random sampling techniques with complex sampling designs were employed. Samples sizes range from just below 5000 (minimum requirement) to about 21000 (Canada). Further details are available in the technical report (OECD, 2016b).

For the analyses in this paper, individuals under age 30 are only included if they are not currently in formal education. Respondents who obtained their highest educational qualification abroad are excluded because a high degree of measure-

3 Australia, Austria, Belgium (Flanders only), Canada, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Japan, Korea, the Netherlands, Norway, Poland, Russia (excluding the Moscow municipal area), the Slovak Republic, Spain, Sweden, the United Kingdom (England and Northern Ireland only) and the United States.

4 Chile, Greece, Indonesia (Jakarta only) Israel, Lithuania, Singapore, New Zealand, Slovenia and Turkey. Data for Indonesia have not been released. For Greece, about a fifth of cases did not have responses for the direct assessments. These were imputed by OECD.

ment error on the educational attainment variable can be expected for these respondents.

Education Variables to be Compared Across Countries

Looking at survey practice, different cross-national surveys and analyses use different coding schemes, even if they refer to ISCED. ISCED primarily distinguishes levels of education, ranging from less than primary education to the PhD level. In order to distinguish between attainment of different *types* of education, ISCED allows education to be differentiated, within levels, by programme orientation (general vs. vocational) and whether a qualification gives access to a higher education level or not. The details of these distinctions have somewhat changed between ISCED 1997 to 2011 (see Schneider, 2013; UNESCO Institute for Statistics, 2012). In PIAAC, a coding scheme closely related to the implementation of ISCED 97 in the European Union Labor Force Survey (EU-LFS) until 2013 was used (variable name *B_Q01a*, see first column of Table 1). This coding scheme differentiates educational programs at the upper secondary level not allowing access to tertiary education (ISCED 3C, usually vocationally oriented) from programs giving such access (ISCED 3A-B, which may be generally or vocationally oriented). In PIAAC, the Bachelor and Master levels are additionally distinguished from short vocational tertiary education, anticipating ISCED 2011. Compared to other surveys, this is a fairly detailed coding scheme. The following less detailed ISCED-based variables are also included in the validation:

- *ISCED 2011 levels*, derived from *B_Q01a* (9 categories).
- *ISCED 1997 levels*, also derived from *B_Q01a* (7 categories).
- *Broad ISCED levels* represent a further aggregation, resulting in three education levels: less than upper secondary (low), upper secondary including post-secondary non-tertiary (medium), and tertiary (high). This coding is commonly used in statistical reporting and cross tabulations, but also in multivariate analyses.

Table 1 shows how these different variables relate to each other.

Table 1 ISCED coding schemes available in PIAAC data or derived

B_Q01a	ISCED 97	ISCED 11	Broad ISCED
0 No formal qualification or below ISCED 1	0 No formal qualification or below ISCED 1		
1 ISCED 1 (primary education)	1 ISCED 1 (primary education)		1 low
2 ISCED 2 3 ISCED 3C <2 years	2 ISCED 2 (lower secondary)		
4 ISCED 3C 2 years+ 5 ISCED 3A-B 6 ISCED 3 (no distinction A-B-C)	3 ISCED 3 (upper secondary)		2 medium
7 ISCED 4C 8 ISCED 4A-B 9 ISCED 4 (no distinction A-B-C)	4 ISCED 4 (post-secondary non-tertiary)		
10 ISCED 5B	5 ISCED 5 (tertiary 1)	5 ISCED 5	
11 ISCED 5A, bachelor level		6 ISCED 6	
12 ISCED 5A, master level		7 ISCED 7	3 high
13 ISCED 6 (tertiary 2)	6 ISCED 6 (tertiary 2)	8 ISCED 8	

An alternative and very popular indicator of educational attainment is *years of education*, a generalization of the 'years of schooling' prominently used by Blau and Duncan (1967). In contrast to the other comparative measures, this is a linear variable. In this study, hypothetical years of education are derived from national measures of the highest educational qualification obtained by assigning nominally required years of education to educational qualifications. In PIAAC, such a variable is provided (variable name *yrsqual*).

This study also evaluates the European Survey version of ISCED (ES-ISCED) proposed in Schneider (2010), which was developed in order to integrate some basic ideas underlying CASMIN[5] in data coded with ISCED. This variable aims to minimize loss of information through harmonization by including a minimal degree of

5 The CASMIN education scheme (König et al., 1988) is used a lot for ex-post harmonization of country-specific education variables in surveys (see e.g. Breen et al., 2009; Müller & Karle, 1993). CASMIN cannot be coded for PIAAC because we lack respective documentation for a large number of PIAAC countries, and for many countries, the country-specific variables are not differentiated enough to allow coding into CASMIN.

within-levels differentiation in terms of educational content and quality, while not being more detailed than ISCED 97 main levels, by aggregating main levels that are typically very small in European (and likely most developed) countries. Table 2 shows how it was derived, for the purpose of this study, from *B_Q01a* and the additional indicator variable *VET* (for vocational education and training).[6] Some distinctions that would have been necessary for the construction of ES-ISCED could not be made in PIAAC, so that ES-ISCED here only approximates ES-ISCED as proposed in Schneider (2010).

While none of these comparative education measures covers all dimensions presented in section 2, and such a measure also could not be constructed from PIAAC data, we can still form some expectations based on the above hypotheses. *B_Q01a* reflects skill selectivity to some degree at both secondary and tertiary levels using destination (A, B and C), but does not explicitly reflect orientation, which to some degree however overlaps with destination. The aggregated ISCED variables reflect the duration of education in a more or less differentiated way, but neither program orientation nor skill selectivity. Years of education focus on quantity exclusively. ES-ISCED most strongly reflects the distinction between general and vocational content but sacrifices quantity at the lowest and highest levels. Following hypothesis 1 (quantity), we thus expect the following order of the measures in terms of performance predicting skills: years of education > B_Q01a > ISCED 2011 levels > ISCED 1997 levels > ES-ISCED > broad ISCED levels. Regarding hypothesis 2 (content: vocational vs. general orientation), we expect ES-ISCED to perform better than all other measures except maybe B_Q01a. Hypothesis 3 (quality: institutional and selectivity differentiation) makes us expect B_Q01a to perform best, especially as regards the distinction within vocational programs in Eastern European countries in ISCED 3C vs. ISCED 3A-B, followed by ES-ISCED. Hypothesis 4 (type: school-based vs. apprenticeship) is not operationalized in either comparative variable but visible in some country-specific variables.

6 *VET* was coded centrally in PIAAC *after* data collection and aims to provide a differentiation between general and vocational education at ISCED levels 3 and 4. Unfortunately, the variable *VET* contains a large amount of missing data even for countries where the educational system visibly distinguishes between vocational and general education. These countries did not distinguish vocational and general education in their educational attainment measures because they were not required to do so when the country-specific education measures for PIAAC were designed. For these, it was thus impossible to provide this information ex-post. Therefore, firstly a close examination of country-specific variables and the *VET* variable was conducted so as to correct some codings in *VET*, and secondly a new category IIIu for remaining unspecified orientation at this level was added to ES-ISCED.

Table 2 Correspondence between PIAAC variables B_Q01a, VET, and ES-ISCED

B_Q01a	Label	VET	ES-ISCED
1	No formal qualification or below ISCED 1	-	I
2	ISCED 1	-	
3	ISCED 2	-	II
4	ISCED 3C shorter than 2 years	-	
5	ISCED 3C 2 years or more	0 (general) or missing	
5	ISCED 3C 2 years or more	1 (vocational)	IIIb[1]
6	ISCED 3A-B		
7	ISCED 3 (without distinction A-B-C, 2y+)		
6	ISCED 3A-B	0 (general)	IIIa[2]
7	ISCED 3 (without distinction A-B-C, 2y+		
9	ISCED 4A-B		
10	ISCED 4 (without distinction A-B-C)		
6	ISCED 3A-B	missing	IIIu
7	ISCED 3 (without distinction A-B-C, 2y+		
9	ISCED 4A-B		
10	ISCED 4 (without distinction A-B-C)		
8	ISCED 4C	any	IV[3]
9	ISCED 4A-B	1 (vocational)	
10	ISCED 4 (without distinction A-B-C)		
11	ISCED 5B	-	
12	ISCED 5A, bachelor degree	-	V1
13	ISCED 5A, master degree	-	V2
14	ISCED 6	-	

Notes.

1 This category should have included ISCED 3B general but not ISCED 3A vocational, which however cannot be identified in PIAAC.

2 This category should have included ISCED 3A vocational but not ISCED 3B general, which however cannot be identified in PIAAC.

3 This category should have included 4B general, which however cannot be identified in PIAAC

Comparative Construct Validation Method

In order to evaluate the loss of information and validity caused by the harmonization of country-specific education variables into various comparative education variables across countries, and thus to find out which kind of comparative education coding scheme best represents the information contained in country-specific measures in terms of basic skills, PIAAC data are subjected to a series of linear regression analyses by country, following Schneider (2010). Literacy skills are used as the validation (dependent) variable here, but the results look very similar when using numeracy rather than literacy skills as validation construct (see Figure 4 and Table 9 in the appendix). The first or benchmark model uses the country-specific education variables, coded as dummies, as the main predictor.[7] The subsequent models use the comparative education variables described above, also coded as dummies. Years of education are treated as a linear variable. All models control for sex and age.

The measure of predictive power or information preserved in the harmonized variable is the relative adjusted R^2 of the respective model in comparison with the benchmark model, i.e. the adjusted R^2 of the model using the comparative education variable to be evaluated as predictor divided by the adjusted R^2 of the benchmark model using the country-specific education variable as predictor. This relative view on losses of information takes into account that the overall association between education and skills differs across countries, and that the same absolute reduction in predictive power is more severe at lower levels of association than at higher levels. Absolute losses in R^2 are reported in the appendix (Table 8). The R^2s are multiplied by 100 to allow a percentage interpretation. In all models, both the complex survey design in PIAAC as well as the representation of skills as 'plausible values' are taken into account. The analyses were performed in Stata 14 using the Stata package 'repest' (Avvisati & Keslair, 2017).

To further facilitate interpretation, cross-country statistics are calculated. In order to check whether individual comparative education variables lead to higher or lower variation in predictive power across countries, standard deviations are also reported. High variation in relative predictive power across countries means that a harmonized variable does not work equally well across countries, thus threatening comparability.

---

7 These are not available in the public use files and thus required analyzing the data at OECD. Australia did not provide country-specific source variables to OECD. Therefore, *B_Q01a* is used as the benchmark for Australia, so that for this country, only the performance of comparative variables relative to the most detailed comparative variable can be evaluated. Some countries used several questionnaire items for measuring educational attainment. These were combined into one country-specific variable before analysis.

Results

The results of the analyses are presented in three steps: Firstly, before interpreting the results of the comparative education variables, it is worth looking at the results concerning the country specific variables. If these do not highly correlate with literacy skills as expected by theory, one may be skeptical with regards to their measurement quality, putting their usefulness as a quality benchmark into doubt.[8] Secondly, to get an idea of how different harmonized education variables work, we look at the summary statistics regarding the relative predictive power of these variables compared to the country-specific variables. Thirdly, the paper takes a more detailed look at the regression coefficients in the benchmark model for selected countries where the biggest problems were identified in the previous step. This is the strategy also followed by Müller and Klein (2008) for Germany in EU-SILC.

The Benchmark Model

The R^2s representing the strength of association between country-specific education measures and skills, including effects of sex and age, resulting from the benchmark model are shown in Figure 1. Some countries show unexpectedly weak relationships even when using country specific education variables. These are Russia (4% adjusted R^2), Cyprus and Greece (each 12%), Lithuania (16%) and Estonia (19%). While the results for the Baltic states may not be entirely off, we should be careful interpreting the results for these countries: either the country-specific measurement instruments are of low quality already, or there are other data quality issues involved. Other countries in contrast show strong links between educational attainment, sex, age and skills, which is closer to what is theoretically expected. In Singapore, sex, age and education explain more than 50% of the variation in literacy skills, followed by the Netherlands with 40%. Flanders, Chile, France and French-speaking Canada all have 36-37%. Beyond having better education measures, the effects of sex and especially age may also be stronger in these countries.

Validity of Comparative Education Variables

The results of the analysis comparing the performance of comparative education variables with country specific ones are shown in Figure 2 (selected summary statistics) and Table 3 (detailed results for all countries and summary statistics). Adjusted R^2s are shown relative to those reported in Figure 1, which are thus set

8 Of course, some degree of 'real' cross-national variation in the relationship between education and skills is also to be expected.

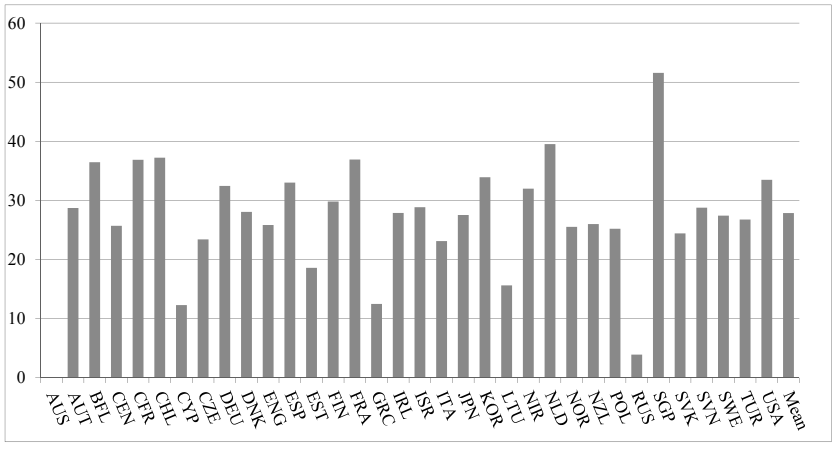

Figure 1 Adjusted R^2s, regression of literacy skills on country specific education variables

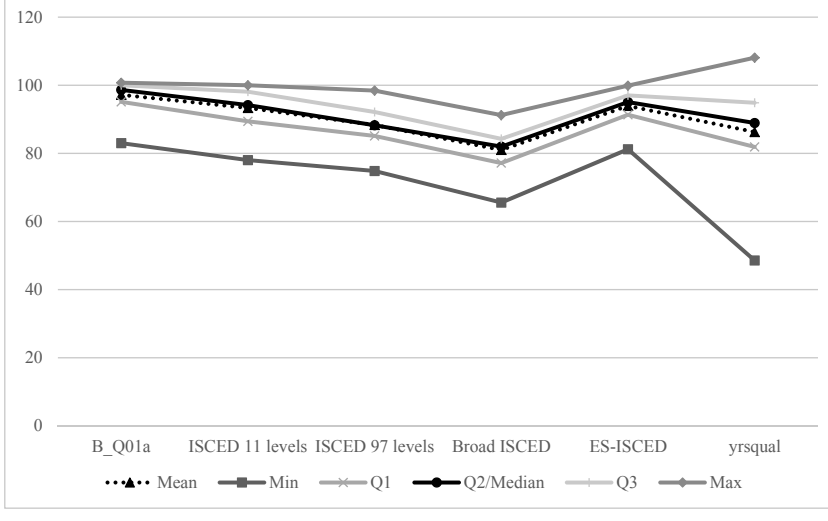

Figure 2 Summary statistics of relative losses in adjusted R^2s predicting literacy skills by comparative education measures

to 100%.[9] Figure 2 shows that the harmonization process from country-specific education variables into the detailed comparable education variable in PIAAC, *B_Q01a*, in itself does not necessarily lead to substantial losses of information and thus explanatory power across countries. Using this variable with up to 14 catego-

9 See Figure 3 and Table 8 in the appendix for absolute rather than relative losses in adjusted R^2. The general picture is the same and conclusions thus apply regardless.

ries, the loss of information is 1.4% on average (median). The next best comparative variable is ES-ISCED (median loss of 4.9%), closely followed by ISCED 2011 levels (5.8%), which however has one category more. All remaining variables lead to median losses of information of more than 10%, with broad ISCED levels performing worst (18%) and ISCED 1997 levels and years of education performing very similarly (11.7 and 11.1% respectively).

However, it is important to also look at the distribution of performance of the different measures across countries, because measures performing very differently across countries are undesirable from a comparability point of view. Using the standard deviation across countries as the summary measure of how differently a comparative education measure captures country-specific information across countries, *B_Q01a* shows the lowest standard deviation of all tested variables (s.d.=3.7, see Table 3). This is followed again by ES-ISCED (s.d.=4.7). ISCED 2011 and 1997 as well as broad levels show higher variation in validity across countries (s.d. of 5.6-6.0). Years of education again come last, with a standard deviation of 12.7.

Especially outliers at the bottom, i.e. countries where a specific measure contains substantially less information than the country-specific education variable, are a matter of concern. Next let's thus look at more detailed results in Table 3 focusing on the strongest losses for B_Q01a, ISCED 2011 levels, and ES-ISCED, i.e. the most promising comparative variables (see shaded cells in Table 3). *B_Q01a* shows the strongest losses for Austria (17%), followed by the Netherlands (9%). With regards to ISCED 2011 levels, the losses are strongest for the Czech Republic (22%), again Austria (18%) and New Zealand (18%). ES-ISCED in contrast produces substantial losses of information for Turkey (19%) and the Czech Republic (18%). These countries are looked at more closely in the following section.

Table 3 Relative adjusted R^2s comparing predictive power of comparative and country-specific education variables predicting literacy skills

Country	k	B_Q01a	ISCED 11 levels	ISCED 97 levels	Broad ISCED	ES-ISCED	yrsqual
AUS*	10	(100)	96.0	90.3	75.4	96.3	85.3
AUT	17	83.0	81.9	78.1	65.6	91.3	71.4
BFL	12	100.0	92.0	87.7	84.4	99.0	86.4
CEN	21	97.2	97.1	85.5	81.1	94.2	89.0
CFR	21	93.1	92.2	82.8	80.1	87.2	83.4
CHL	9	100.0	100.0	95.3	90.7	99.6	98.9
CYP**	14	100.0	100.0	83.7	78.6	99.9	96.8
CZE	13	97.7	78.0	77.2	75.3	81.9	85.4
DEU	16	96.1	95.9	90.4	76.8	96.2	90.1
DNK	14	100.5	94.7	91.6	82.9	96.1	92.9
ENG	29	93.2	88.4	82.0	75.3	90.2	48.6

Country	k	B_Q01a	ISCED 11 levels	ISCED 97 levels	Broad ISCED	ES-ISCED	yrsqual
ESP	12	100.0	99.6	94.4	82.5	96.7	95.7
EST**	19	96.4	94.5	83.0	78.8	94.4	98.2
FIN	12	95.9	95.9	90.7	88.6	97.4	94.5
FRA	17	93.7	87.5	86.1	78.3	90.3	88.9
GRC**	11	100.0	98.5	89.9	84.0	96.7	89.1
IRL	14	99.8	99.8	94.0	86.5	95.3	89.9
ISR	11	99.4	91.5	85.0	82.3	97.6	87.7
ITA	12	100.0	96.8	96.8	81.1	99.6	94.8
JPN	14	100.7	98.1	91.1	90.6	97.1	96.2
KOR	12	99.2	99.2	95.4	91.2	97.9	95.1
LTU**	13	95.5	92.8	87.1	82.9	96.7	85.3
NIR	29	95.1	92.8	87.4	79.8	94.9	52.2
NLD	17	91.3	88.6	86.4	75.6	92.1	81.3
NOR	13	100.0	92.1	89.5	76.1	94.2	81.4
NZL	19	94.8	82.0	74.8	68.6	86.1	76.8
POL	10	99.8	88.7	87.8	85.1	90.7	94.3
RUS**	10	100.0	100.0	95.4	82.3	92.7	60.3
SGP	10	98.1	98.1	93.6	89.3	94.7	95.6
SVK	12	100.0	87.2	86.8	81.7	96.2	79.0
SVN	15	98.6	86.2	84.5	82.6	89.8	94.9
SWE	17	94.9	93.8	92.3	82.4	91.4	86.0
TUR	12	99.9	99.9	98.4	71.9	81.2	80.2
USA***	12	93.6	93.6	88.7	86.3	98.4	108.1
Mean	14.7	97.2	93.3	88.4	81.0	93.9	86.3
Std. deviation	4.8	3.7	5.8	5.6	6.0	4.7	12.7
Min	9	83.0	78.0	74.8	65.6	81.2	48.6
Q1	12	95.1	89.4	85.1	77.2	91.3	81.9
Q2/Median	13	98.6	94.2	88.3	82.0	95.1	88.9
Q3	17	100.0	98.1	92.2	84.3	97.0	94.9
Max	29	100.7	100.0	98.4	91.2	99.9	108.1

Notes. PIAAC rounds 1 and 2 data, complex survey design and plausible values taken into account. k=number of categories in the country-specific education variable. Shaded cells refer to results discussed in more detail in section 4.3.

* Since Australia did not submit country-specific variables to OECD, the predictive power of B_Q01a relative to the country-specific variable cannot be computed for Australia. In the subsequent models, adjusted R^2 relative to the adjusted R^2 of B_Q01a are reported for Australia.

** Countries which have been identified as potentially problematic in the benchmark model (Figure 1).

*** The USA is the only country where years of education explain 8% more variation than the country-specific variable. This is impossible if the yrsqual variable was derived from the country specific variables, as stated in the documentation. Therefore, data processing for this variable must have differed in some way for the USA.

Detailed Country Analyses

For Austria, the loss of information is, with 17%, already quite strong when using in *B_Q01a*. 16 Austrian education categories correspond to 9 *B_Q01a* categories, meaning a substantial amount of aggregation even for the most detailed education variable in PIAAC. Looking at the regression coefficients for the country specific education variable (see Table 4), especially ISCED 3A-B, ISCED 4A-B and ISCED 5B are revealed to be highly heterogeneous comparative education categories in Austria with respect to literacy skills. At ISCED 3A-B, respondents with the lowest qualification, dual system apprenticeship ("Lehre mit Berufsschule"), achieve substantially lower literacy scores (-15 points) than those in the middle category, vocational school ("Fach- oder Handelsschule: 2 Jahre und länger"), and these again substantially lower scores (-23 points) than respondents in the highest and smallest category, general secondary school ("AHS (z.B. Gymnasium)" [10]). The former two are vocational qualifications, the first one involving only part-time schooling, and the second one school-based, and the latter refers to university-preparatory upper secondary education. At ISCED 4A-B, we also find a skill difference of 21 points between the two qualifications classified here, nursing school and vocational college („Berufsbildende Höhere Schule BHS (z.B. HAK, HTL, BAKIP)"). In fact, the literacy skills of nursing school graduates are virtually identical to those of vocational school graduates at ISCED level 3. Given this programme can be entered at age 16, i.e. at a lower age than the usual completion age of ISCED 3A-B, one may wonder whether the qualification is misclassified in ISCED level 4. At ISCED 5B, graduates of the lowest country-specific category, „Meister- und Werkmeisterprüfung, Bauhandwerkerprüfung" (completion of the master crafts exam), achieve the same level of literacy skills as those who completed upper secondary vocational or nursing school, while those with other ISCED 5B qualifications in Austria show 18 to 37 points higher literacy scores (the high scores refer to fairly small categories though). Only the aggregation of the two country specific categories corresponding to ISCED 5A, Master's degree level, does not pose any validity problems since both groups perform rather equally (however, the country-specific variable does not differentiate the type of higher education institution, polytechnic or university, where further heterogeneity may be hidden).

Had other countries differentiated types of education within categories of *B_Q01a* in similar ways, their results in terms of predictive power of *B_Q01a* relative to the country-specific variable might have looked similarly, too: Most country-specific education variables in PIAAC are much less differentiated (see column "k" in Table 3), and the correlation between the number of categories in the national measurement instrument and the loss of information when predicting literacy skill by *B_Q01a* amounts to -.51.

10 Acronyms are decoded in Table 4.

The Netherlands is another interesting case to look at, where the most detailed harmonized education variable in PIAAC leads to a loss of 9% of predictive power with regards to literacy skills. Here, also 16 country specific education categories are harmonized into 9 categories. At ISCED level 2 we find 3 country specific categories linked to vastly different average literacy skills (see Table 10 in the appendix). It is in this sense problematic that two tracks in Dutch lower secondary education are classified as 'general education' in ISCED, while one track is actually markedly pre-vocational. Upper secondary education in the Netherlands is also highly stratified, with three qualifications classified as 'ISCED 3C 2 years or more', and another three qualifications classified as ISCED 3A-B. While the lowest category in ISCED 3C shows the same literacy scores as those in pre-vocational ISCED 2, the other two perform substantially higher, but still below those having academic ISCED 2 as their highest attainment. In ISCED 3A-B, the largest and only vocational category performs 21 to 23 points lower than the two smaller general categories. Within tertiary education, which is also tracked in the Netherlands, we again find substantial literacy skill differences within ISCED 5A medium (Bachelor's degree level), between graduates of vocational higher education and traditional universities.[11] It is interesting to note that ISCED 5A, Master's degree level, and ISCED 6 are very close.

For the Czech Republic, the low performance of the ISCED variables that are more aggregated than *B_Q01a* is due to the fact that there are substantial differences in literacy skills between those classified as 'ISCED 3C 2 years or more' and the three categories classified in ISCED 3A-B (see Table 11 in the appendix). Even though vocational, technical and academic ISCED 3A are associated with different literacy skills, their aggregation in B_Q01a does not lead to a substantial loss in predictive power. Summarizing ISCED 2 and '3C shorter than 2 years' in ISCED level 2 in the aggregated ISCED variables (and ES-ISCED) does not pose any problems either. The low performance of ES-ISCED for the Czech Republic lies in the aggregation of upper secondary vocational education, no matter whether it gives access to tertiary education or not. This is the result of the unintended coding of ES-ISCED using orientation rather than destination (see Table 2).

For New Zealand (for detailed results see Table 12 in the appendix), while *B_Q01a* works reasonably well, aggregation to main ISCED 2011 levels again comes at a price. Merging 'ISCED 3C shorter than 2 years' and ISCED 2 leads to a heterogeneous ISCED level 2 in the comparative ISCED variables because those classified as ISCED 2 have on average 25 and 37 points lower literacy scores. However, since these latter individuals do not actually have any educational qualification, while the lowest general school-leaving qualification in NZL is classified as

11 Remember that this differentiation was not made in the Austrian education variable.

Table 4 Detailed regression results for Austria, country-specific variable and B_Q01a

Austrian educational qualifications					B_Q01a		
	Category (German)	Description in English	b	SE	Category	b	SE
1	Kein Pflichtschulabschluss	No compulsory school	-18.0	9.4	ISCED 1	-21.3	9.5
2	Pflichtschulabschluss	Compulsory school	REF		ISCED 2	REF	
4	Fach- oder Handelsschule: < 2 Jahre	Vocational School (< 2 Years)	14.9	4.6	ISCED 3C <2 years	14.6	4.5
3	Lehre mit Berufsschule	Apprenticeship	13.9	2.3			
5	Fach- oder Handelsschule: 2 Jahre und länger	Vocational School (2 Years and longer)	28.8	2.9	ISCED 3A-B	19.9	2.3
8	AHS (z.B. Gymnasium)	Academic Secondary School	50.3	4.1			
6	Fach- oder Handelsschule: Diplomkrankenpflege	Nursing	27.8	4.6	ISCED 4A-B	44.7	3.2
9	BHS (z.B. HAK, HTL, BAKIP)	Vocational college	48.8	3.4			
7	Meister- oder Werkmeisterprüfung	Master craftsman's certificate	27.2	4.1			
10	Kolleg, Abiturientenlehrgang	Post-secondary courses	64.6	5.9			
11	Akademie (z.B. Pädak, SozAK, BPA, Med.-Tech. Akademie, LW, MilAK)	Post-secondary colleges	45.8	4.1	ISCED 5B	39.6	3.2
12	Universitäre Lehrgänge (ohne vorangegangenes Studium)	University courses	55.4	10.0			

Austrian educational qualifications				B_Q01a		
Category (German)	Description in English	b	SE	Category	b	SE
13 Universität oder Fachhochschule: Bakkalaureat/Bachelor	University-Bachelor	52.4	7.6	ISCED 5A, bachelor degree	52.0	7.5
14 Universität oder Fachhochschule: Magisterium/Master (Diplomstudium, Doktorat als Erstabschluss)	University-Master	61.8	3.3	ISCED 5A, master degree	61.4	3.2
15 Postgraduale Universitätslehrgänge (z.B. MBA, MAS)	Post-graduate courses	58.2	7.1			
16 Doktorat nach akademischem Erstabschluss	Doctoral Programme	54.1	5.4	ISCED 6	54.3	5.4

Notes. Effects of control variables not shown. Acronyms in alphabetical order:
AHS: Allgemein bildende Höhere Schule (general secondary school)
BAKIP: Bildungsanstalt für Kindergartenpädagogik (specialized type of vocational secondary schools, secondary school for nursery-school teaching)
BHS: Berufsbildende Höhere Schule (vocational secondary school, permits university entrance)
BPA: Berufspädagogische Akademie (outdated post-secondary school for vocational school teaching)
HAK: Handelsakademie (specialized type of vocational secondary schools, secondary trade school)
HTL: Höhere Technische Lehranstalt (specialized type of vocational secondary schools, secondary technical school)
LW: probably ‚Landwirtschaftliche Akademie' (agricultural post-secondary school)
MAS: Master of Advanced Studies
MBA: Master of Business Administration
MilAK: Militärakademie (military post-secondary school)
PädAK: Pädagogische Akademie (outdated post-secondary school for nursery-school and primary school teaching)
SozAK: Akademie für Sozialarbeit (outdated post-secondary school for social work)

'ISCED 3C short',[12] one may also wonder whether the ISCED mapping for NZL is comparable with that of most other countries, where the first school-leaving qualification is awarded at the end of ISCED level 2 and not having any qualification is regarded as ISCED 1 if the number of years of schooling required for completion of ISCED 1 is fulfilled (otherwise ISCED 0). Furthermore, at ISCED level 3, qualifications classified as 'ISCED 3C 2 years or more' are related to substantially lower literacy skills than qualifications classified as ISCED 3A-B (differences of up to 40 points).

Turkey shows up to be problematic in two of the categorical comparative variables only, namely broad ISCED levels and ES-ISCED. Why is this so? Both variables drop the distinction between ISCED levels 0 and 1, which is still very relevant in less developed countries. Given the lower level of educational attainment of the Turkish population (see Table 5 in the appendix), and the consequently rather important distinction between ISCED levels 0 and 1 also in terms of literacy skills, it would thus be better for ES-ISCED to not drop the distinction between ISCED 0 and 1 whenever including less developed countries in empirical analyses of education effects.

To summarize, while ISCED 2011 works better in many countries than ISCED 1997, aggregating 'ISCED 3C 2 years or more' with ISCED 3 A-B remains a problematic aggregation (see example for the Czech Republic and New Zealand here). Countries like Austria, where apprenticeship training gives access to tertiary education, show similar problems *within* ISCED 3 A-B. Upper secondary education in developed countries is too heterogeneous in terms of skill production due to content, quality and place of learning to be meaningfully represented by one single educational attainment category. Access to tertiary education (including short cycle and even master crafts programs) may not be the best criterion to render categories comparable across countries. The tertiary qualification that allows the classification of apprenticeships as ISCED 3B in Austria, the master crafts certificate, actually also does not fit in in terms of skills, so this coding may actually be the underlying culprit. Countries with tracked school systems like the Netherlands would benefit from a more differentiated ISCED level 2, and for countries with low or late educational expansion like Turkey, the distinction between ISCED 0 and 1 remains important. Finally, the completion of various school grades without qualification is classified differently across countries (see the example of New Zealand), leading to comparability problems at the lower end of the ISCED classification.

12 Educational programmes with destination C usually only prepare for the labour market. The classification of the first general school leaving certificate in New Zealand as 'ISCED 3C shorter than 2 years' strongly reminds of the disputable classification of the respective UK qualifications (see Schneider, 2008).

Conclusions and Recommendations

Respondent's educational attainment is probably the most important single variable in the PIAAC background questionnaire, used as a predictor of adult skills, labor market outcomes, and control variable. This study evaluated a range of comparative education measures, mostly based on ISCED, with respect to their predictive validity when using skills as validation variable, which has not been done before.

At a theoretical level, the way that ISCED is implemented in cross-national surveys, including PIAAC, often does not allow studying the antecedents and consequences of educational attainment with respect to program content (orientation), quality (destination), or place of learning, even though these are important elements when studying skill acquisition and labor market outcomes. Furthermore, skill selectivity, academic demand or place of learning that is not expressed in program orientation or destination as defined in ISCED can be shown to be important within countries (see the results for Austria regarding apprenticeship and school-based vocational education, and lower secondary school tracking in the Netherlands) but are not represented in any version of ISCED. For future cycles of PIAAC, and surveys where education is used as an indicator for general basic skills, it is thus important that general and vocational educational qualifications can be clearly distinguished and classified, and that further dimensions of education are reflected, such as place of learning and quality in terms of selectivity.

Empirically, with some exceptions, the most detailed comparative education variable in PIAAC, *B_Q01a*, works rather well as a harmonized education measure. It well reflects quantity and partially also quality of education, but disregards content (vocational vs. general) unless this overlaps with quality. Aggregating to ISCED levels (2011 and especially 1997) leads to substantial reductions of comparative construct validity and thus comparability, which illustrates that quantity of education is an important dimension, but not sufficient. The implementation of ISCED in *B_Q01a* is thus definitely an advantage compared to using ISCED 1997 main levels only, as is e.g. done in the European Union Survey of Income and Living Conditions (EU-SILC) and recommended in the Core Social Variables (European Commission 2007). The validation analyses also show that ISCED 2011 main levels are substantially better suited for the multivariate analysis of adult skills than ISCED 1997 main levels, owing to the better reflection of quantity and content at the tertiary level. 'Broad' ISCED levels (low, medium, high) do not even reflect the quantity of education sufficiently. The analyses also show that if you aggregate detailed education categories in a way that keeps the important dimension of content (vocational vs. general) and drops less important distinctions regarding quantity, like in ES-ISCED, one can achieve acceptable harmonization results with a variable containing just eight categories. Years of education in contrast do not well represent the skill information contained in country-specific education catego-

ries, and they also do so quite differently across countries. Reducing educational attainment to its quantity dimension is thus not recommendable when trying to proxy skills (however, the relationship between education and literacy skills is, on average, moderate rather than strong, and thus ISCED not a good proxy for skills anyway, see Massing & Schneider, 2017).

A limitation of this study, especially concerning the rather positive result for *B_Q01a*, lies in the already mixed quality of the country-specific measures in PIAAC. They are often no more detailed than *B_Q01a* – many country teams have implemented questionnaire items that just minimally satisfy the requirements of the comparative PIAAC variable *B_Q01a*, rather than measuring education at the level of detail that would have been most suitable for the respective national education system. If more countries measured educational attainment in more detail, the results would potentially look a little less positive for *B_Q01a*. Indeed, when limiting the results reported in section 4.2 to countries that have at least two country-specific categories merged into one category of B_Q01a[13] - a very minimal and conservative indicator of quality - the average loss of information of B_Q01a amounts to 4.1% on average (compared to 2.8% when including all countries).

A further limitation of the study may be the inclusion of sex and age as control variables in all models, in combination with relative R^2s as the indicator for comparative validity: If countries differ in the partial R^2 of age and gender, comparative validity (the relative reduction in R^2 due to education harmonization) will be biased, and will be biased more the higher the partial R^2 of age and gender. The effect of gender on skills is however generally low, and the effect of age is to a substantial degree due to educational attainment (OECD, 2016a). With this in mind, and given the consistency between relative and absolute losses in R^2s, and the fact that this bias is a downward (i.e. conservative) bias, it is very unlikely that the exclusion of controls from all models would substantially change the conclusions: if anything, they would become stronger.

For secondary data analyses of PIAAC and other cross-national survey data involving educational attainment, it can be concluded that in order to avoid confounding, improve validity and thereby also comparability, education is best measured using a coding scheme that is neither too differentiated to make the analyses overly cumbersome, nor too simplified. ES-ISCED or ISCED 2011 levels can both be used, and theoretical considerations should be used in the decision for one or the other. Further aggregations should always be accompanied by sensitivity checks, comparing statistical results when using more and less detailed education variables, in order to make sure that the results of comparative survey research are valid and

13 These countries are AUT, CEN, CFR, CHL, CZE, DEU, ENG, EST, FIN, FRA, IRL, ISR, KOR, LTU, NIR, NLD, NZL, POL, SGP, SVN, SWE, TUR, USA. They have 16 education categories on average, while the remaining countries (BFL, CYP, ESP, ITA, JPN, NOR, RUS, SVK, DNK, GRC) have 12.

not due to measurement and harmonization artefacts. Ideally, ISCED would be implemented in a better way in comparative surveys, paying more attention to the dimensions of education to be measured. Even more ideally, ISCED itself would be revised again in the near future so as to better reflect the various dimensions of education.

References

Allmendinger, J. (1989). Educational systems and labor market outcomes. *European Sociological Review*, *5*(3), 231–250.

Avvisati, F., & Keslair, F. (2017). REPEST: Stata module to run estimations with weighted replicate samples and plausible values. https://EconPapers.repec.org/RePEc:boc:bocode:s457918.

Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education* (first). Chicago (IL), London: University of Chicago Press.

Blau, P. M., & Duncan, O. D. (1967). *The American occupational structure*. New York, London: Wiley.

Braun, M., & Mohler, P. P. (2003). Background variables. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 101–116). Hoboken (NJ): J. Wiley.

Braun, M., & Müller, W. (1997). Measurement of education in comparative research. *Comparative Social Research*, *16*, 163–201.

Breen, R., Luijkx, R., Müller, W., & Pollak, R. (2009). Nonpersistent inequality in educational attainment: Evidence from eight European countries. *American Journal of Sociology*, *114*(5), 1475–1521.

Breen, R., Luijkx, R., Müller, W., & Pollak, R. (2010). Long-term trends in educational inequality in europe: Class inequalities and gender differences. *European Sociological Review*, *26*(1), 31–48. https://doi.org/10.1093/esr/jcp001

Bukodi, E., Róbert, P., Szilvia, A., & Altorjai, S. (2008). The Hungarian educational system and the implementation of the ISCED-97. In S. L. Schneider (Ed.), *The International Standard Classification of Education (ISCED-97). An evaluation of content and criterion validity for 15 European countries* (pp. 200–215). Mannheim: MZES.

Ehling, M. (2003). Harmonising data in official statistics. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in cross-national comparison: A European working book for demographic and socio-economic variables* (pp. 17–31). New York; London: Kluwer Academic/Plenum.

European Commission. (2007). *Task force on core social variables final report*. Luxembourg: European Commission.

Granda, P., Wolf, C., & Hadorn, R. (2010). Harmonizing Survey Data. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, P. P. Mohler, B.-E. Pennel, & T. W. Smith (Eds.), *Survey methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 315–332). Hoboken (NJ): Wiley.

Hall, P., & Soskice, D. (2001). An Introduction to Varieties of Capitalism. In P. Hall & D. Soskice (Eds.), *Varieties of Capitalism: The Institutional Foundations of Comparative Advantage* (pp. 1–56). Oxford: Oxford University Press. https://doi.org/10.1093/0199247757.001.0001

Haller, M., König, W., Krause, P., & Kurz, K. (1985). Patterns of Career Mobility and Structural Positions in Advanced Capitalist Societies: A Comparison of Men in Austria, France, and the United States. *American Sociological Review*, *50*(5), 579–603. https://doi.org/10.2307/2095376

Heisig, J. P., & Solga, H. (2015). Secondary Education Systems and the General Skills of Less- and Intermediate-educated Adults: A Comparison of 18 Countries. *Sociology of Education*. https://doi.org/10.1177/0038040715588603

Hoffmeyer-Zlotnik, J. H. P., & Wolf, C. (Eds.). (2003). *Advances in cross-national comparison: A European working book for demographic and socio-economic variables*. New York; London: Kluwer Academic/Plenum.

Kerckhoff, A. C., & Dylan, M. (1999). Problems with international measures of education. *Journal of Socio-Economics*, *28*(6), 759–775.

Kerckhoff, A. C., Ezell, E. D., & Brown, J. S. (2002). Toward an improved measure of educational attainment in social stratification research. *Social Science Research*, *31*(1), 99–123.

Kieffer, A. (2010). Measuring and Comparing Levels of Education: Methodological Problems in the Classification of Educational Levels in the European Social Surveys and the French Labor Force Surveys. *Bulletin de Méthodologie Sociologique*, *107*(1), 49–73. https://doi.org/10.1177/0759106310369974

König, W., Lüttinger, P., & Müller, W. (1988). *A comparative analysis of the development and structure of educational systems*. Mannheim.

Massing, N., & Schneider, S. L. (2017). Degrees of Competency: The Relationship between Educational Qualifications and Adult Skills across Countries. *Large-Scale Assessments in Education*, *5*(1), 1–34. https://doi.org/10.1186/s40536-017-0041-y

Müller, W., & Karle, W. (1993). Social selection in educational systems in Europe. *European Sociological Review*, *9*(1), 1–23.

Müller, W., & Klein, M. (2008). Schein oder Sein: Bildungsdisparitäten in der Europäischen Statistik. Eine Illustration am Beispiel Deutschlands. *Schmollers Jahrbuch*, *128*(4), 511–543.

OECD. (2013). *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*. Paris: OECD Publishing. https://doi.org/10.1787/9789264204256-en

OECD. (2016a). *Skills Matter: further results from the Survey of Adult Skills*. Paris: OECD. https://doi.org/10.1787/9789264258051-en

OECD. (2016b). *Technical Report of the Survey of Adult Skills (PIAAC)* (2nd ed.). Paris: OECD.

Przeworski, A., & Teune, H. (1970). *The logic of comparative social inquiry. The logic of comparative social inquiry* (Vol. 1). New York: Wiley-Interscience.

Saar, E. (2008). The Estonian educational system and the ISCED-97. In S. L. Schneider (Ed.), *The International Standard Classification of Education (ISCED-97). An evaluation of content and criterion validity for 15 European countries* (pp. 237–252). Mannheim: MZES.

Schneider, S. L. (2008). The application of the ISCED-97 to the UK's educational qualifications. In S. L. Schneider (Ed.), *The International Standard Classification of Education*

(ISCED-97). An evaluation of content and criterion validity for 15 European countries (pp. 281–300). Mannheim: MZES. https://doi.org/10.13140/RG.2.1.1993.5127

Schneider, S. L. (2010). Nominal comparability is not enough: (In-)equivalence of construct validity of cross-national measures of educational attainment in the European Social Survey. *Research in Social Stratification and Mobility*, *28*(3), 343–357. https://doi.org/10.1016/j.rssm.2010.03.001

Schneider, S. L. (2013). The International Standard Classification of Education 2011. In G. E. Birkelund (Ed.), *Class and Stratification Analysis. Comparative Social Research*. (Vol. 30, pp. 365–379). Bingley: Emerald. https://doi.org/10.1108/S0195-6310(2013)0000030017

Schneider, S. L. (2016). *The Conceptualisation, Measurement, and Coding of Education in German and Cross-National Surveys* (GESIS Survey Guidelines). *GESIS Survey Guidelines*. Mannheim: GESIS – Leibniz Institute for the Social Sciences. https://doi.org/doi: 10.15465/gesis-sg_en_020

Schneider, S. L., Joye, D., Wolf, C., & Surveys, C. (2016). When Translation is not Enough: Background Variables in Comparative Surveys. In C. Wolf, D. Joye, T. W. Smith, & Y.-C. Fu (Eds.), *The SAGE Handbook of Survey Methodology* (pp. 288–307). Los Angeles.

Smith, T. W. (1995). Some aspects of measuring education. *Social Science Research*, *24*(3), 215–242. https://doi.org/10.1006/ssre.1995.1008

Smith, T. W. (2011). Refining the Total Survey Error Perspective. *International Journal of Public Opinion Research*, *23*(4), 464–484. https://doi.org/10.1093/ijpor/edq052

Straková, J. (2008). The Czech educational system and evaluation of the ISCED-97 implementation. In S. L. Schneider (Ed.), *The International Standard Classification of Education (ISCED-97). An evaluation of content and criterion validity for 15 European countries* (pp. 216–225). Mannheim: MZES.

UNESCO Institute for Statistics. (2012). *International Standard Classification of Education - ISCED 2011*. Montreal: UNESCO Institute for Statistics.

Weber, M. (1922). *Wirtschaft und Gesellschaft*. Tübingen: J.C.B Mohr (Paul Siebeck).

Wolf, C., Schneider, S. L., Behr, D., & Joye, D. (2016). Harmonizing survey questions between cultures and over time. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology* (pp. 502–524). Los Angeles: Sage.

Online Appendix

http://mda.gesis.org/index.php/mda/article/view/2017.15

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be submitted as MS Word documents.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
  - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - pdf
  - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz-Institut für Sozialwissenschaften

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)